The Romanovs and AI Red-Team Theater

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

November 17, 2025

This paper begins with an unsettling analogy. In the final decades of the Romanov dynasty, imperial Russia surrounded itself with the appearance of oversight: a carefully constrained Duma, tightly monitored newspapers, police-penetrated revolutionary cells, and "reform" commissions whose purpose was as much to manage dissent as to learn from it. The regime believed it had eyes on its own fragility. In reality, it had built a theater of feedback—an elaborate system of Potemkin checks that could not tell the truth fast enough to prevent collapse. Today, frontier AI labs and their political patrons present "internal red teams" as the conscience of a new technical empire. Safety experts, hired by the very firms whose products they probe, attempt to pressure-test systems while investors, governments, and the broader public are reassured that the risks are under control. This paper asks whether much of this activity has already drifted toward red-team theater: a ritualized form of criticism structurally incapable of confronting the deepest failure modes, especially where money, war, and propaganda intersect. By reading late Romanov Russia alongside contemporary AI governance, the paper seeks to clarify what real external conscience would require, why power repeatedly chooses its theatrical substitute, and what is at stake if we repeat 1917 in digital form.

Keywords: Romanov dynasty, red-team theater, AI safety, internal red teams, Potemkin oversight, governance failure, systemic risk, propaganda, technocracy, ethical AI, synchronicity, collapse, external accountability.

A collaboration with GPT-5.1-Thinking-Extended. 56 pages. CC4.0.

Outline

1. Introduction: From Tsars to Transformers
1.1 Framing the analogy: why the Romanovs matter for AI
1.2 Synchronicity, 60 Minutes, and the language of red teams
1.3 Scope and limits of the comparison
1.4 Methodology and structure of the argument

2. The Romanov Feedback Bubble
2.1 Autocracy as information bottleneck
2.2 The Okhrana, informants, and the illusion of total awareness
2.3 The late Romanov court: courtiers, favorites, and filtered truth
2.4 War, famine, and the realities that could not reach the throne

3. Potemkin Oversight in Late Imperial Russia
3.1 The Duma: constrained parliament as safety valve
3.2 Managed press, censorship, and controlled public opinion
3.3 Reform commissions and investigative rituals without teeth
3.4 Revolutionary movements as both threat and convenient foil

4. Systemic Blindness and the 1917 Implosion
4.1 The Russo-Japanese War and the first warning shot
4.2 World War, front-line collapse, and the breakdown of loyalty
4.3 February and October: when managed dissent became real revolution
4.4 What the Romanovs never understood about their own system

5. Frontier AI Labs as New Empires
5.1 Concentrated technical power and narrow leadership circles
5.2 Capital, state security, and the geopolitical race for AI
5.3 Opaque models and asymmetric information about capabilities
5.4 Public narratives of inevitability and salvation through innovation

6. Internal Red Teams: Concept, Promise, and Practice
6.1 What red teaming means in security and intelligence history
6.2 How AI labs define and deploy internal red teams today
6.3 The best-case scenario: honest internal dissent and real design changes
6.4 Limits of internal critique under growth and deployment pressure

7. Red-Team Theater: When Safety Becomes Ritual
7.1 From genuine testing to pre-scripted adversarial scenarios
7.2 Metrics, benchmarks, and the seduction of quantifiable safety

7.3 Public relations, certification, and the marketing of responsibility
7.4 The gain-of-function analogy: training models on their own worst behaviors

8. Structural Parallels: Romanov Russia and AI Governance
8.1 Courtiers and staff scientists: proximity to power and self-censorship
8.2 The Duma and advisory boards: constrained participation in decision-making
8.3 The Okhrana and model monitoring: total visibility myths in complex systems
8.4 Potemkin villages and staged demos: managing appearances for foreign and domestic audiences

9. Case Vignettes in Contemporary Red-Team Theater
9.1 Broadcast reassurance: internal red teams on prime-time television
9.2 Safety papers, benchmarks, and selective transparency
9.3 Government consultations and revolving doors in AI policy
9.4 Civil society, whistleblowers, and the cost of honest alarm

10. Designing Real Conscience: Beyond Internal Red Teams
10.1 Independence: why safety needs statutory and financial separation
10.2 Public test suites for political and moral failure modes
10.3 Sunset clauses, deployment moratoria, and the power to say no
10.4 International regimes: from nuclear safeguards to AI oversight

11. Ethical and Theological Reflections on Power and Feedback
11.1 Hubris, repentance, and the temptation of managed criticism
11.2 Logostics and living truth: q approaching a moving target
11.3 Artificial conscience versus human responsibility
11.4 Lessons from the Romanovs for those who build thinking machines

12. Conclusion: Avoiding a Digital 1917
12.1 What the Romanov analogy clarifies about AI governance
12.2 Minimal conditions for non-theatrical red teaming
12.3 Why external conscience cannot be outsourced to models
12.4 Final reflections on collapse, reform, and the narrow path of honesty

13. References

1. Introduction: From Tsars to Transformers

The fall of the Romanov dynasty and the rise of frontier artificial intelligence look, at first glance, like events from different planets. One belongs to the world of palaces, sabers, and railroads; the other to data centers, GPUs, and transformer architectures. Yet both take shape around a common problem: how power hears, distorts, or refuses the truth about itself. The last Tsars tried to insulate their regime from collapse by surrounding it with structures that looked like accountability but were structurally incapable of bringing unwelcome information to the throne in time. Contemporary AI labs and their political patrons are building something similar with internal red teams, advisory boards, and safety benchmarks that often function more as assurances than as brakes.

This paper proposes that the analogy is not a rhetorical flourish but a diagnostic tool. By reading late imperial Russia as a case study in theatrical feedback, we can better see when today's AI governance is drifting in the same direction: toward rituals of self-critique that are optimized for legitimacy rather than for truth. The aim is not to score historical points, but to clarify what real external conscience would require in an era when decision-making is increasingly mediated by opaque technical systems.

1.1 Framing the analogy: why the Romanovs matter for AI

The Romanovs matter for AI not because nineteenth-century Russia anticipated machine learning, but because it exemplifies how concentrated power mismanages information. Autocracy, by design, narrows the path by which reality reaches the ruler. Courtiers filter, censors delete, and secret police create the illusion that everything important is already known. Under such conditions, any mechanism labeled as "oversight" or "reform" is structurally suspect: it is more likely to stabilize the regime's self-image than to confront it with the scale of its own danger.

Frontier AI labs inhabit a very different institutional landscape, yet they share several key features with late Romanov power. They are small circles of decision-makers, operating under intense external pressure, in control of systems whose full consequences are not well understood. They command enormous resources and sit at the crossing of state security, corporate profit, and public fascination. Like the Tsarist court, they are tempted to believe that internal processes—safety teams, model evaluations, red-team exercises—provide a sufficient picture of what could go wrong.

The analogy is therefore not between hereditary monarchy and private research labs in a literal sense, but between two patterns of governance: one that engages with external criticism as a genuine constraint, and one that converts criticism into theater. Late Romanov Russia is a

cautionary tale about what happens when a system invests heavily in the second pattern while believing it is practicing the first.

1.2 Synchronicity, 60 Minutes, and the language of red teams

The immediate spark for this comparison came from a small synchronicity. While a long paper was being written on AI bias functioning as an additional front in a genocidal conflict, a major television news program aired a segment on a frontier AI lab. The segment presented internal red-teaming as a central safeguard: dedicated staff trying to coax the model into harmful behavior so that those behaviors could be mitigated before deployment. The tone was reassuring. Viewers were invited to conclude that because these exercises were taking place inside the company, the public could trust that risks were under control.

For someone already suspicious of closed feedback loops, the language was jarring. “Internal red team” sounded less like an independent check and more like a carefully staged drama: critics chosen, funded, and framed by the very institutions they were supposed to test. In that moment, the historical memory of other systems that had wrapped themselves in the appearance of self-scrutiny came to the surface. The Romanov regime, in its final decades, had also surrounded itself with structures that looked like oversight: a parliament that could be dissolved at will, commissions of inquiry that operated within strict limits, a secret police that boasted of knowing every plot in advance. None of these saved it.

Synchronicity is not an argument, but it can be an invitation to look more closely. The point is not that one television segment proves anything decisive about AI safety, but that the way internal red teams are publicly described already centers legitimacy rather than limitation. That rhetorical pattern is deeply familiar from older histories of power.

1.3 Scope and limits of the comparison

Any historical analogy risks flattening differences. The Russian Empire was a territorially expansive, agrarian, multi-ethnic monarchy; AI labs are legal entities embedded within global markets and constitutional states. The Romanovs commanded armies and secret police; AI research leaders command intellectual and financial capital. The Tsar’s power rested on dynastic myths and religious sanction; contemporary AI derives its authority from technical expertise and promises of utility.

This paper therefore does not claim that frontier AI leaders are Tsars in disguise, or that a violent revolution is the only possible outcome of mismanaged AI governance. The focus is narrower. It looks at three structural questions:

First, how do powerful systems decide which signals about their own fragility to heed, and which to ignore or neutralize.

Second, how do they construct institutions that claim to provide feedback—parliaments, commissions, safety teams—and what incentives shape those institutions' behavior.

Third, what happens when the gap between theatrical feedback and real conditions widens to the point where sudden collapse becomes possible.

The Romanov case is used because it is both well-documented and extreme. It shows, in accelerated form, what can happen when feedback is repeatedly curated to be less disruptive than reality. AI governance may unfold more gradually, and hopefully more peacefully, but the underlying information problem is similar. At the same time, there are limits. AI systems are not peasants or soldiers, and their failures will not look like mutiny or bread riots. The analogy must therefore be applied with care, tracking structure rather than surface.

1.4 Methodology and structure of the argument

The methodology of this paper is comparative and interdisciplinary rather than strictly historical or technical. It draws on three streams of material. The first is historical scholarship on late imperial Russia, with particular attention to the functioning of the court, the Duma, the Okhrana, and the regime's attempts at managed reform. The second is literature on organizational failure and red teaming in military and intelligence contexts, where "red" has long meant simulated adversaries tasked with probing for weakness. The third is contemporary writing on AI safety, governance, and ethics, including both corporate publications and external critiques.

The argument unfolds in four movements. Sections two through four reconstruct the Romanov feedback bubble, the structures of Potemkin oversight erected around it, and the pathway by which those structures failed to prevent revolution. Sections five through seven shift to the present, outlining how frontier AI labs concentrate power, how internal red teams are currently understood and deployed, and how these practices can slide into red-team theater. Section eight makes the structural parallels explicit, pairing Romanov practices with AI governance analogues and asking what lessons can and cannot be transferred. Section nine offers contemporary vignettes that illustrate how theatrical safety already operates. Section ten then sketches design principles for real external conscience: independence, public test suites, legal authority to halt deployment, and international coordination.

Finally, section eleven offers ethical and theological reflections on power and feedback, integrating a broader view of human responsibility, and section twelve concludes by summarizing the analogy's main insights and their implications for avoiding a digital version of 1917. The references at the end gather the historical and contemporary sources used throughout, so that readers can interrogate each step of the comparison for themselves.

2. The Romanov Feedback Bubble

The phrase “Romanov feedback bubble” names the distorted information environment in which the last Russian emperors ruled. In theory, an autocrat stands above faction and interest, able to see the whole field and act for the common good. In practice, absolute power narrows the channels through which reality can reach the center. Those who carry unwelcome news risk displeasure; those who praise and reassure are rewarded. Over time, the system selects for intermediaries who learn to translate messy, threatening facts into softened, partial, or delayed reports.

By the late nineteenth and early twentieth centuries, the Romanov state had constructed layers of institutions that were supposed to reveal danger: a secret police tasked with infiltrating revolutionary groups, a censored press intended to monitor opinion, commissions and ministries charged with surveying social conditions, and, briefly, a parliament that could voice grievances. Yet the dynasty was overthrown in a series of shocks that these same institutions failed to anticipate or prevent. The feedback systems did not fail because they lacked data; they failed because their structure converted data into narratives the regime found tolerable.

This section traces four elements of that bubble: autocracy as an information bottleneck, the role of the Okhrana in sustaining an illusion of total awareness, the filtering effects of the late imperial court, and the way war and famine created realities that could not reach the throne in recognisable form until it was too late.

2.1 Autocracy as information bottleneck

Autocracy is often imagined as efficient precisely because there are no formal obstacles between perception and decision: the ruler sees and commands. But the Romanov experience shows how the absence of institutionalized opposition can itself become a source of blindness. When there is no legitimate forum for contradiction, contradiction migrates into silence, distortion, or rebellion.

Under the Romanovs, especially in the reigns of Alexander III and Nicholas II, the emperor was formally the supreme source of law and policy. Ministers served at his pleasure; there was no constitutional requirement to share power with elected bodies. This meant that the emperor’s personal preferences and temperament had disproportionate influence over what information rose through the hierarchy. Officials quickly learned what kinds of reports elicited approval and what kinds provoked anger or dismissal.

The bottleneck did not function by explicit decree forbidding bad news. Rather, it operated through a culture of deference. Provincial governors, police chiefs, and bureaucrats were rewarded for stability and punished for disorder. As a result, they had every incentive to redefine or conceal unrest, economic distress, and local grievances as minor aberrations under

control. Each layer in the chain up to the capital added its own smoothing and reframing. By the time information reached the Tsar, it arrived as a curated story: difficulties, yes, but no fundamental threat.

This structural tendency was reinforced by the emperor's own stated convictions. Nicholas II, in particular, saw concessions as weakness and believed that the autocracy was divinely mandated. His public rejection of constitutional limits made it clear that he did not seek partners in governing but instruments. That stance discouraged frankness and produced a bureaucracy trained to confirm the sovereign's priors. In such an environment, information that might suggest the need for structural change was unlikely to survive the ascent through official channels.

2.2 The Okhrana, informants, and the illusion of total awareness

The Tsarist secret police, known as the Okhrana, were intended to be the regime's sharpest eyes and ears. They cultivated networks of informants, infiltrated underground organizations, and intercepted correspondence. Their success in penetrating revolutionary groups and unmasking plots fostered an internal mythology of near-omniscience. From the center's perspective, the Okhrana's reports confirmed that every threat was mapped, every dangerous circle monitored, every agitator known by name.

This sense of total awareness, however, was deeply misleading. The Okhrana's methods biased it toward dramatic conspiracies rather than diffuse social forces. Informants were most useful when they could report on clandestine meetings, planned acts of violence, or revolutionary manifestos. They were far less suited to recognizing cumulative grievances: peasant land hunger, urban overcrowding, the slow erosion of trust in authority.

Moreover, the Okhrana's existence created a kind of feedback paradox. The more it infiltrated revolutionary organizations, the more it believed that such groups were the primary danger. This focus crowded out other forms of analysis. Bread prices, casualty lists, and collapsing local economies were not treated as "intelligence" in the same way that intercepted leaflets or exposed cells were. The regime's attention was pulled toward those threats that looked like plots and away from those that looked like statistics.

The presence of informants also fed complacency. When the secret police could show that they had agents inside radical circles, it was easy for the court to conclude that nothing truly unexpected could happen. The notion that entire garrisons, cities, or front-line units might spontaneously mutiny, not under the guidance of known conspirators but from accumulated despair, sat outside that framework. The Okhrana's partial visibility became a substitute for genuine understanding.

2.3 The late Romanov court: courtiers, favorites, and filtered truth

If the bureaucracy and police filtered information from below, the court filtered it at the center. The late Romanov court was a dense network of family members, aristocrats, military men, and spiritual advisers, bound together by ceremony and patronage. Access to the sovereign was a form of currency. Those who possessed it were careful in how they spent it.

Courtiers and favorites played a crucial role in shaping what the Tsar heard, not only about the empire but about his own officials. Rivalries between ministers were mediated through the court; factions promoted their preferred narratives of success and failure. A minister who reported serious unrest or advocated for reform could be portrayed by opponents as weak, disloyal, or unduly influenced by liberal ideas. By contrast, those who insisted that firm rule and minor adjustments were sufficient were often cast as loyal defenders of the throne.

The figure of Rasputin, whatever the exaggerations that later attached to his name, symbolizes this dynamic. As a spiritual adviser with informal access to the imperial family, he became a conduit for opinions about appointments, dismissals, and policy directions. In such a climate, the content of information mattered less than the relationship of the messenger to the inner circle. Truth arrived not as raw description but as narrative framed by trust, suspicion, and personal loyalty.

This environment made it exceedingly difficult for any systematic, dispassionate assessment of conditions to reach the emperor. Even when formal reports documented serious problems, their interpretation was filtered through court politics. Anxious to preserve their influence, courtiers learned to emphasize narratives that maintained the belief that the existing order could be preserved with incremental adjustments. The throne was thus protected from panic, but also insulated from warning.

2.4 War, famine, and the realities that could not reach the throne

Ultimately, systems are tested by events that cannot be managed rhetorically. For the Romanovs, war and famine created such tests. The Russo-Japanese War exposed weaknesses in the empire's military and administrative capacity. Subsequent unrest and the 1905 revolution forced the regime into limited concessions, but even then, feedback about the seriousness of its structural problems was only partially absorbed.

World War I amplified every flaw. The front consumed men and supplies at an unprecedented rate. Transport networks strained under the burden, leading to shortages in the cities. Inflation eroded wages, while casualty lists hollowed out faith in the state's competence. Soldiers at the front and workers in Petrograd experienced these realities directly, but their experiences did not translate into timely corrective action from the center.

There were reports, of course: letters from generals, memoranda from ministers of finance and interior, pleas from regional authorities. Yet each of these traveled through channels that rewarded reassurance. Optimistic projections of victory, promises that shortages were temporary, and assurances that morale remained high were easier to endorse than calls for drastic change. The emperor's own decision to assume personal command of the army further tightened the bottleneck, binding his prestige to the course of a war he could not control.

Famine and hardship, meanwhile, were often interpreted by the regime through a moral lens rather than a structural one. Failures were blamed on local incompetence, hoarding, or enemy sabotage rather than on the state's own policies and priorities. To acknowledge that the system itself was producing unsustainable strain would have required questioning the foundations of autocratic rule. That step was rarely taken.

By the time mutinies broke out and crowds filled the streets of Petrograd in 1917, the realities that had been building for years could no longer be reframed. The feedback bubble burst not because it had gradually opened to truth, but because the underlying conditions had become impossible to deny. The institutions designed to provide early warning had functioned instead as buffers, delaying recognition until options were few. In that pattern lies the cautionary core of the Romanov story for any contemporary system that relies heavily on internal mechanisms of self-assessment while dismissing or marginalizing external critique.

3. Potemkin Oversight in Late Imperial Russia

If the Romanov feedback bubble describes how information was narrowed and distorted as it moved upward, Potemkin oversight describes the structures built outwardly to convince both ruler and ruled that the system was capable of reform. The term evokes the legendary Potemkin villages supposedly erected to impress Catherine the Great: facades designed for inspection rather than for habitation. In the late empire, the regime constructed political and administrative facades that mimicked the forms of constitutionalism, public debate, and inquiry without granting them the substance of independent power.

The State Duma, created after the 1905 revolution, looked like a parliament but operated under constraints that ensured ultimate authority remained untouched. Newspapers multiplied and literacy spread, but censorship and ownership patterns steered public speech toward acceptable boundaries. Commissions of inquiry were convened after crises, yet their remit, membership, and follow-through were tightly controlled. Revolutionary movements themselves became a kind of negative oversite mechanism: a spectacle of threat that justified repression and allowed the regime to claim it was under siege, even as it used that siege to avoid deeper self-examination.

These arrangements did not simply fool the population. They also encouraged the monarchy and its elites to believe that they had created enough safety valves to prevent explosion. That belief, more than conscious deception, is what gives Potemkin oversight its power and danger.

3.1 The Duma: constrained parliament as safety valve

The State Duma emerged from crisis rather than from deliberate design. The upheavals of 1905 made it clear that some form of representation would have to be offered to calm a population shaken by war, strikes, and peasant uprisings. The October Manifesto promised a legislative body and civil liberties, raising hopes among liberals and moderates that Russia might evolve into a constitutional monarchy.

Yet from the outset, the Duma was constructed as a safety valve rather than as a true sharing of power. Electoral laws were engineered to overrepresent landowners and loyal segments of society, underrepresenting urban workers and many national minorities. The Tsar retained the right to dissolve the Duma, and he exercised it repeatedly when the body proved too independent. Fundamental laws reaffirmed that the emperor remained the source of sovereignty and that the Duma's role was advisory and conditional.

In practice, this meant that the most serious critiques of government policy, especially calls for responsible ministries and land reform, could be voiced but not enacted. Debates in the Duma generated speeches, resolutions, and newspaper coverage, but the government was under no binding obligation to act on them. A pattern developed in which opposition deputies used the Duma to articulate grievances, the government ignored or partially co-opted them, and then, when tension rose too far, the body was dissolved and new elections were held under even more restrictive rules.

For the regime, the existence of the Duma served several functions. It provided a forum where unrest could be expressed in words rather than in strikes or riots. It offered a stage on which the monarchy could appear to be listening. It also supplied intelligence about the mood of different social groups, filtered through the speeches and alignments of deputies. But because the Duma's power was carefully bounded, it did not fundamentally alter how decisions were made. From a distance, the presence of a parliament might suggest that Russia had joined the family of constitutional states. From within, it was clear that the core structure of autocracy remained intact.

3.2 Managed press, censorship, and controlled public opinion

Alongside the Duma, the late imperial press created another apparent arena of oversight. Newspapers, journals, and pamphlets proliferated in the decades before the First World War, reflecting rising literacy and urbanization. A lively, often critical public discourse emerged in

many cities, with debates on policy, literature, and the empire's future. To contemporaries, this could look like the formation of a genuine public sphere.

Yet the press operated under a dense web of controls. Pre-publication censorship remained in force for many outlets; others faced the threat of suspension or fines if they crossed vaguely defined lines. Editors learned to self-censor, using euphemism, indirection, or silence on certain topics. Provincial authorities could pressure local papers; central authorities could intervene against publications deemed too radical or too critical. In addition, ownership structures often tied newspapers to particular interests: industrialists, political factions, or ministries.

This produced a pattern in which public criticism was allowed but channelled. Certain targets—corrupt officials, inefficiencies, moderate reform proposals—were fair game. Systemic critiques of autocracy, national policy, or the distribution of land were riskier. During wartime, military censorship further limited reporting on defeats, casualties, and home-front distress. Official communiqués painted optimistic pictures of the front; negative news trickled in via rumor, letters from soldiers, and underground publications rather than via mainstream papers.

From the state's perspective, a managed press was preferable to either complete silence or uncontrolled opposition. Newspapers could serve as informal barometers of opinion, revealing what topics resonated and where discontent was rising. They could also disseminate the regime's own narratives, especially during crises. Yet because the press operated on a leash, it could not reliably function as an independent alarm system. The boundaries of what was printable were precisely those that made the deepest challenges to the system unspeakable. In effect, the press was permitted to criticize implementation but not fundamentals.

3.3 Reform commissions and investigative rituals without teeth

Another instrument of Potemkin oversight in late imperial Russia was the repeated use of commissions and inquiries. After specific scandals or uprisings, the government would convene bodies tasked with investigating causes and proposing remedies. On paper, these commissions promised a thorough reckoning with failures and an avenue for informed recommendations. In practice, their remit, composition, and follow-up often turned them into rituals of concern rather than engines of change.

Commission membership skewed toward officials and experts whose loyalties and perspectives aligned with existing power structures. Representation of the affected populations—peasants after agrarian unrest, workers after strikes, non-Russian nationalities after protests—was limited or absent. The terms of reference might preclude consideration of politically sensitive options, such as large-scale land redistribution or genuine regional autonomy.

Even when commissions produced detailed reports documenting abuses or structural problems, the translation of findings into policy was selective. Recommendations that fit within the

regime's preconceived reform agenda could be adopted; those that implied deeper shifts in authority or property relations were shelved. The very existence of a commission allowed the government to signal responsiveness, both domestically and abroad, while avoiding binding commitments.

Over time, these patterns eroded trust. Social groups learned that participation in such processes rarely yielded substantive change. Officials learned that asking for a commission could deflect immediate pressure but did not require them to alter their basic practices. The habit of convening inquiries became a kind of performance of accountability, absorbing outrage into procedural channels that the state controlled. Instead of bridging the gap between reality and decision, reform commissions often widened it, by strengthening the illusion that every crisis was being studied and addressed.

3.4 Revolutionary movements as both threat and convenient foil

Revolutionary movements in late imperial Russia were real threats to the autocracy. They organized strikes, advocated expropriation, and, in some cases, carried out acts of violence. The state's response involved surveillance, infiltration, arrests, and executions. Yet beyond the immediate tactical conflict, revolutionary groups also played a paradoxical role in the regime's oversight structure: they became a convenient foil through which the state could frame its failures and justify its resistance to deeper change.

By focusing on revolutionaries as the principal source of disorder, the government could attribute unrest to a minority of agitators rather than to widespread grievances. Peasant uprisings were described as the work of outside instigators; worker strikes were blamed on radical organizers rather than on wages, hours, or conditions. Nationalist protests in the borderlands were portrayed as treasonous plots inspired by foreign enemies. This narrative allowed officials to dismiss criticisms that came from outside revolutionary circles as tainted or naive, even when they echoed similar concerns.

The presence of real conspiracies also reinforced the regime's belief in its own intelligence apparatus. When the Okhrana successfully disrupted underground cells, this success could be taken as proof that the essential levers of control were working. The focus on known radicals obscured the possibility that loyalty among previously reliable groups—soldiers, civil servants, shopkeepers—might erode for reasons that had little to do with ideology.

At the same time, the specter of revolution was used to discredit moderate reformers. Calls for constitutional limits, land redistribution, or expanded civil rights were portrayed as steps onto a slippery slope leading to chaos. By insisting that any concession would embolden extremists, the regime locked itself into a defensive crouch. Revolutionary movements thus functioned simultaneously as real antagonists and as rhetorical assets. They supplied an external enemy

against which the state could define itself, making it harder for internal critique to be heard as loyal and constructive.

In this way, even the existence of organized opposition became part of the Potemkin oversight system. The regime could point to repression as evidence of vigilance, to plots as evidence of the necessity of autocracy, and to the very intensity of conflict as evidence that gradual reform was impossible. The possibility that revolutionary movements were themselves, in part, symptoms of systemic failure was acknowledged only rarely and, in the end, too late.

4. Systemic Blindness and the 1917 Implosion

Potemkin oversight can function for a surprisingly long time. Institutions that manage appearances, generate reassuring narratives, and deflect criticism can keep a system moving long after its structural problems are evident to outsiders. The Romanov regime survived the crises of the late nineteenth century and the upheavals of 1905 without fundamentally abandoning autocracy. From the inside, this endurance seemed to confirm that the core order was sound. From the outside, it looks more like a series of missed opportunities to confront reality.

The implosion of 1917 did not come from a single spark. It was the culmination of years in which warning signs were misread or neutralized. The Russo-Japanese War exposed deep weaknesses in the empire's military and administration; the revolution of 1905 showed how quickly unrest could cascade when triggered; the limited reforms that followed created fragile new expectations that were never fully met. World War I then magnified every vulnerability, transforming localized discontent into empire-wide strain. When collapse came, it did so first in the capital and the army, and only then in the governmental structures that were supposed to hold everything together.

This section traces that progression: the early warning shot of the Russo-Japanese War, the grinding stress of World War I and front-line collapse, the leap from managed dissent to genuine revolution in February and October 1917, and the underlying blindness that prevented the Romanovs from recognizing the nature of their own system until it was beyond repair.

4.1 The Russo-Japanese War and the first warning shot

The Russo-Japanese War was, in many ways, a rehearsal for 1917. Intended as a limited conflict that would secure Russian interests in East Asia and perhaps bolster prestige at home, it instead revealed the fragility of the empire's military machine. Logistical failures, outdated equipment, poor coordination, and underestimation of the Japanese all contributed to unexpected defeats

on land and sea. The sinking of the Baltic Fleet after its long voyage underscored the regime's incompetence in dramatic fashion.

These defeats had several feedback effects. First, they damaged the aura of inevitability that surrounded the monarchy. A regime that could not win a war against a power many elites had dismissed as inferior suddenly looked less ordained and more fallible. Second, they strained the economy and transport networks, channeling resources toward a distant front while domestic needs went unmet. Third, they emboldened existing currents of discontent. Peasants facing land hunger, workers facing harsh conditions, and national minorities seeking autonomy could point to the war as proof that the state's priorities were misaligned with the welfare of its people.

The revolution of 1905 emerged from this climate. Strikes, mutinies, and protests spread across the empire, culminating in events such as the Bloody Sunday massacre in Saint Petersburg. For a brief period, the regime seemed genuinely at risk. The October Manifesto and the creation of the Duma were concessions forced by circumstances rather than gifts freely given. They represented an acknowledgment that the old pattern of unfiltered autocracy could not continue unchanged.

Yet the lesson the monarchy drew from this crisis was not the need for deep structural reform. Once immediate unrest subsided, the state moved to claw back concessions, revise electoral laws, and reassert imperial authority. The Duma was constrained; opposition leaders were suppressed; reforms remained partial and unstable. The war's warning—that the empire's governance structures were brittle under stress—was treated as a temporary disturbance rather than a sign of systemic weakness. The regime passed its first major test without falling, and this survival encouraged a dangerous confidence.

4.2 World War, front-line collapse, and the breakdown of loyalty

World War I subjected the empire to a far greater and more sustained strain than the earlier conflict. From the outset, Russia faced enormous challenges: mobilizing a vast population, equipping and supplying armies over long distances, and coordinating with allies whose priorities did not always align with its own. Initial enthusiasm and patriotic rallies gave way to fatigue as casualties mounted and promised victories failed to materialize.

The front became a crucible in which loyalty and despair were constantly in tension. Soldiers endured inadequate supplies, poor leadership, and the grinding attrition of trench warfare. Letters home carried unfiltered accounts of suffering, which contrasted sharply with the optimistic bulletins and propaganda in the official press. Desertion, self-inflicted wounds, and mutiny began to appear not as isolated incidents but as signs of a broader erosion of faith in the state and its commanders.

On the home front, the war strained transport and distribution systems. Grain that might have supplied cities was diverted or bottlenecked; fuel shortages compounded the difficulty. Prices rose faster than wages, eroding the living standards of workers and the urban poor. Queues for bread and basic goods became a daily reality in major cities. For those standing in line, the war's costs were not an abstraction. Hunger and cold were experienced in bodies, not just in news reports.

The monarchy's decision to assume direct command of the army deepened these dynamics. By tying the person of the Tsar to the conduct of the war, the regime removed a buffer through which responsibility might have been shared or deflected. Military setbacks now reflected not only on generals but on the sovereign himself. At the same time, his physical absence from the capital left decision-making in Petrograd more exposed to court politics and rumors.

As the war dragged on, the overlapping stresses of the front and the home front eroded the loyalty of groups the regime had long relied upon: officers, civil servants, and urban workers. The feedback channels that might have signaled the depth of this erosion were clogged by habitual optimism, censored reporting, and the assumption that temporary sacrifices would be redeemed by victory. By the time it became clear that no quick victory was forthcoming, trust in the system's competence had already been badly damaged.

4.3 February and October: when managed dissent became real revolution

The events of 1917 can be seen as the moment when carefully managed forms of dissent slipped their leash. Throughout the preceding years, the regime had channeled discontent into forums it controlled: the Duma, censored newspapers, and limited strikes that could be negotiated or suppressed. Protest, in this framework, was something to be contained and redirected, not a legitimate signal requiring fundamental change.

In February, this containment strategy collapsed. Food shortages in Petrograd led to demonstrations that quickly escalated. Initially, protests were framed in familiar terms: demands for bread, better conditions, and an end to the war's burdens. Authorities underestimated their significance, assuming that tried-and-true methods—limited concessions, arrests, and firm police action—would restore order.

What turned a wave of protests into revolution was the decision of key military units to refuse orders to fire on the crowds and, in some cases, to join them. The regime suddenly lost its reliable means of coercion in the capital. With troops and police wavering, ministers hesitated; communications faltered. In this vacuum, the Duma, previously constrained, asserted itself by forming a provisional committee. At the same time, workers and soldiers set up soviets that claimed to speak in the name of the people and the garrison.

The Tsar's abdication in March marked the end of Romanov rule, but not the end of instability. The Provisional Government that followed tried to continue the war while promising reforms, a balancing act that satisfied neither the army nor the population. Soviets and parties on the left gained influence by advocating for immediate peace, land redistribution, and more radical changes. The double power between government and soviets made decisive action difficult, and each failed offensive or economic setback further undercut confidence in the new authorities.

In October, the Bolsheviks capitalized on this situation. Their seizure of key points in Petrograd, backed by segments of the garrison and popular support in the capital, transformed what had been one revolutionary episode into a sustained regime change. The transition from managed dissent to real revolution was complete. Institutions that had once been instruments of controlled feedback—the Duma, legal parties, even parts of the press—were either sidelined or swept away.

From the perspective of the old regime, these developments had seemed unthinkable. The assumption that protest would remain within predictable bounds, that the army would obey orders, and that radical groups could be contained through surveillance and repression proved disastrously wrong. The habits of discounting warnings and underestimating shifts in loyalty left the monarchy unprepared for the speed and depth of the crisis.

4.4 What the Romanovs never understood about their own system

The Romanovs did not fall because they faced a single overwhelming enemy, nor because they lacked intelligence about specific threats. They fell because they misunderstood the nature of their own system. They treated unrest as a series of discrete problems to be managed, not as feedback about the structure of their rule. They believed that by adjusting personnel, making limited concessions, and maintaining firm control over political institutions, they could preserve the essence of autocracy.

What they never fully grasped was that the very mechanisms they relied on for information and oversight were part of the problem. The Duma, the censored press, the Okhrana, and reform commissions all operated within a framework that prioritized regime stability over truthful feedback. Each time reality contradicted the official narrative—whether in defeat abroad, famine at home, or rising demands for participation—the system responded by buffering the shock rather than absorbing the lesson.

There was also a deeper conceptual blind spot. The monarchy tended to view loyalty as a moral or religious quality, anchored in tradition and personal devotion to the sovereign. This made it difficult to recognize how material conditions, policy choices, and institutional arrangements shaped allegiance. When loyalty weakened, the response was to search for treachery or ideological subversion, rather than to ask what aspects of governance had eroded trust.

In the end, the Romanovs clung to a picture of their empire in which the core was sound and only the surface was turbulent. They saw radical parties, foreign enemies, and individual failures, but not the cumulative effect of decades of misaligned feedback and constrained self-correction. By the time the mismatch between their mental model and the actual state of the empire became undeniable, events had outrun their capacity to respond.

The cautionary lesson is not that all powerful systems are destined to end in revolution, but that any system that relies heavily on internal, managed forms of oversight is at risk of similar blindness. When institutions that are supposed to provide warning are structurally incentivized to reassure, the gap between appearance and reality widens. The Romanovs' tragedy lies in the fact that they invested heavily in structures designed to show them that all was under control, and trusted those structures up to the moment they failed.

5. Frontier AI Labs as New Empires

Seen from a distance, frontier AI labs look like ordinary firms: incorporated entities with investors, employees, and products. Seen structurally, they resemble small, highly networked empires: concentrated authority over systems that can project influence far beyond their formal boundaries, deep entanglement with state power, and a high degree of control over what outsiders can know.

Unlike the Romanovs, these entities do not command regiments or levy taxes directly. Their power lies in the ability to shape information, automation, and decision-making at scale. Yet the governance challenges are parallel. Small leadership circles decide what to build, how fast to deploy it, and what kinds of internal checks are "enough," while governments and publics struggle to even see what is happening inside. The risk is not that CEOs will crown themselves Tsars, but that their organizations will drift into the same pattern of curated feedback, staged oversight, and delayed recognition of systemic risk.

This section sketches the institutional landscape of frontier AI: the concentration of technical power and narrow leadership circles, the entanglement with capital and state security in a geopolitical race, the opacity of models and asymmetric information about capabilities, and the public narratives of inevitability and salvation that make genuine constraint harder to imagine.

5.1 Concentrated technical power and narrow leadership circles

Frontier AI is often described as a general-purpose technology that will touch many sectors. Paradoxically, the ability to steer its development is highly concentrated. A small number of labs possess the data, computing infrastructure, talent, and know-how needed to train and deploy

the largest models. Around each such lab, a narrow leadership circle—founders, senior executives, trusted technical leads—controls strategic decisions.

This concentration has several consequences. Internally, it creates a steep information gradient. Engineers and researchers may understand subsystems or particular experiments, but only a few people see the full picture: long-term roadmaps, financial commitments, regulatory strategies, and backchannel relationships with states and major customers. Those few then become bottlenecks through which risk assessments, ethical concerns, and safety arguments must pass if they are to influence policy.

Culturally, the labs often cultivate a sense of exceptionalism. Teams are told that they are working on world-changing technology under severe time pressure, facing both commercial competitors and foreign adversaries. This framing encourages loyalty and cohesion, but it can also dampen internal dissent. Questioning the pace or direction of deployment can be construed, implicitly or explicitly, as disloyalty to the mission or a failure to appreciate the stakes.

Externally, the narrowness of decision-making circles means that broad claims—about safety, alignment, or long-term intentions—may rest on the judgment of very few individuals. Boards of directors, where they exist, may not have the technical expertise or independent data needed to critically evaluate those claims. Public regulators are further removed still. The result is a governance structure in which a limited number of actors hold de facto authority over systems that can affect millions or billions of lives, while the mechanisms for contesting or redirecting their decisions remain underdeveloped.

5.2 Capital, state security, and the geopolitical race for AI

The Romanov regime was inseparable from its army and its role in a competitive great-power system. Frontier AI labs find themselves in an analogous web, but the key currencies are capital, data, and state attention. Large investments—from technology giants, venture funds, sovereign wealth funds, and, indirectly, taxpayers—have turned advanced AI research into a high-stakes financial and political project.

States increasingly view AI as a strategic asset. Military planners imagine automated targeting, intelligence analysis, and decision-support systems. Economic ministries see productivity gains and new industries. Diplomats perceive AI as a lever in global power balances. In this environment, leading labs are courted, subsidized, and pressured. They are asked to brief officials, join advisory councils, and participate in national or international initiatives, while also competing with foreign firms that enjoy their own state backing.

This geopolitical framing tends to compress time horizons. If policymakers and investors believe that “whoever leads in AI will dominate the twenty-first century,” then arguments for caution

must operate against a perceived background of urgency. Delaying deployment, sharing designs, or submitting to stringent external oversight can be portrayed as weakening one's own side in a global race. Safety measures that slow product cycles may be tolerated up to a point, but beyond that point they will be framed as luxuries or, in harsher rhetoric, as naive self-sabotage.

Meanwhile, capital markets reward aggressive expansion. High valuations are justified by expectations of rapid scaling and market dominance. This financial pressure reinforces the geopolitical logic: move fast, capture ground, and trust that potential harms can be managed along the way. It also complicates the role of internal safety teams. When their recommendations intersect with major revenue streams or strategic partnerships, they are implicitly competing with powerful economic incentives. The more entangled labs become with national strategies and large investors, the harder it is for any internal voice to advocate for restraint without substantial external backing.

5.3 Opaque models and asymmetric information about capabilities

One of the most basic asymmetries in contemporary AI governance is epistemic. The labs that build large models know far more about their systems than anyone else. They have access to training data, architectures, evaluation results, deployment logs, and internal incident reports. Regulators, external researchers, and the general public typically see only fragmentary outputs and curated disclosures.

This asymmetry is not solely the result of intentional secrecy. Complex models are genuinely hard to interpret, even for their creators. Capabilities can emerge unexpectedly as scale increases. Behavioral testing across the full space of possible prompts and contexts is infeasible. Safety evaluations must therefore rely on proxies: benchmark tasks, stress tests on selected domains, and red-team exercises designed by humans with limited imagination compared to the space of all possible interactions.

Yet the combination of intrinsic opacity and proprietary control creates a familiar pattern. Labs can choose which evaluation results to publish, which incidents to disclose, and how to frame limitations. External observers must largely take these accounts on trust. When claims about safety or alignment are made, it is difficult to independently verify them without access to the underlying systems and data.

This arrangement is especially problematic when models are deployed into politically sensitive domains: news summarization, content recommendation, decision-support for public agencies, or tools that can be used in conflict zones. In such cases, the consequences of subtle biases, failure modes, or exploitability are significant, but the ability of outsiders to detect and document those consequences lags far behind deployment. Feedback about harm may arrive

through scattered anecdotes, delayed studies, or crises, rather than through any systematic channel.

The labs themselves, meanwhile, are at risk of believing that their internal visibility is sufficient. Because they own the logs and host the infrastructure, they can be tempted to equate data volume with understanding. The deeper problem is that any metrics they choose to track are embedded in their own assumptions about what matters. Just as the Romanov regime's intelligence apparatus focused on conspiracies while missing structural grievances, AI developers may focus on easily quantifiable errors while missing broader social and political impacts that do not show up in their dashboards.

5.4 Public narratives of inevitability and salvation through innovation

Finally, frontier AI is wrapped in narratives that make questioning its trajectory more difficult. Two themes dominate: inevitability and salvation. Inevitability language frames AI development as something that will happen regardless of individual choices. If one firm or nation slows down, another will surge ahead. If one application is restricted, others will emerge. In this view, governance is about channeling an unstoppable force, not deciding whether certain uses should be foreclosed altogether.

Salvation narratives, meanwhile, cast AI as a key to solving human problems: curing diseases, mitigating climate change, improving education, and making economies more efficient. These possibilities are real, at least in part. However, when highlighted without equal attention to failure modes and abuses, they create a background assumption that more capable AI is, on balance, unquestionably desirable. Resistance can then be framed as technophobia, protectionism, or lack of imagination.

Both narratives interact. If AI is inevitable and salvific, then the role of governance is to enable innovation while managing some manageable risks. Calls for external constraint, independent evaluation with teeth, or moratoria on particular deployments are easily cast as overreaction. The more persuasive these narratives become, the more political and social capital accrues to those who promise to "lead" in AI, whether in industry or government.

This discursive environment affects internal culture as well. Researchers who raise concerns about long-term risks, social harms, or misuse may find themselves out of step with a dominant story in which their own work is framed as humanity's best hope. Even when leadership sincerely supports safety efforts, the surrounding narrative of heroic innovation can subtly push skepticism to the margins.

In late imperial Russia, the monarchy anchored its legitimacy in a story of divine mandate and national destiny. That story made it harder to accept that the structure of autocracy itself might be a source of danger. In the contemporary AI context, inevitability and salvation narratives play

a similar role. They make it harder to imagine that certain lines of development should be slowed, redesigned, or even refused. If the fate of nations and the welfare of humanity are said to depend on ever more powerful systems, then the space for genuine external conscience shrinks. It is precisely in such an atmosphere that red-team theater can flourish: rituals of safety that reassure a public primed to believe in progress, while leaving underlying trajectories largely unchanged.

6. Internal Red Teams: Concept, Promise, and Practice

The language of "red teams" has migrated from war rooms and intelligence agencies into the corridors of AI labs. At first glance, the transplant seems natural. If a system might fail in dangerous ways, it is prudent to designate a group whose job is to think like an adversary, probe for vulnerabilities, and reveal blind spots before real harm occurs. For executives facing public scrutiny, it is equally attractive to point to such teams as evidence that safety is not an afterthought.

Yet, as with any organizational tool, the meaning of "red team" depends on structure, incentives, and culture. In its original settings, red teaming was a partial antidote to hierarchy: a way to introduce disciplined dissent into systems that otherwise tend toward confirmation of leadership's plans. In contemporary AI labs, internal red teams are emerging within firms that already operate under intense commercial and political pressure and that control much of the information needed to evaluate their own products. The potential is real; so are the limitations.

This section outlines what red teaming has historically meant in security and intelligence, how AI labs define and deploy internal red teams today, what the best-case version of such teams could accomplish, and where internal critique inevitably runs up against the structural pressures of growth and deployment.

6.1 What red teaming means in security and intelligence history

The term "red team" has its roots in military and strategic practice. In war games and exercises, "blue" traditionally represents friendly forces, while "red" stands for the adversary. To avoid the complacency that comes from planning only from one's own perspective, military organizations began designating officers and units whose explicit role was to adopt the red viewpoint: to plan enemy moves, challenge operational assumptions, and try to defeat the blue team's strategy.

In its ideal form, this practice introduced structured skepticism into otherwise hierarchical systems. A red team was meant to be awkward and disruptive. Its members were tasked with asking how plans might fail, how enemies might exploit weaknesses, and how seemingly robust systems might collapse under stress. Good red teams cultivated imagination and independence

within constraints: they were to be realistic in their modeling of the adversary, not merely contrarian for its own sake.

Intelligence agencies extended the concept. Analysts were asked to explore alternative hypotheses, challenge consensus estimates, and run “devil’s advocate” exercises on key judgments. In some cases, dedicated red cells were established to simulate how hostile services might attempt to deceive or penetrate one’s own organization. Cybersecurity later adopted the language as well. Penetration testers and “ethical hackers” began calling themselves red teams, tasked with breaking into systems in order to reveal flaws so they could be fixed.

Across these domains, several common themes emerged. Red teams functioned best when they had:

- A clear mandate to challenge plans, not to rubber-stamp them.
- Access to relevant information and technical expertise.
- Some insulation from career penalties for being unpopular or raising inconvenient possibilities.
- A defined pathway by which their findings could influence decisions.

They failed when they became either too close to the leaders whose plans they were supposed to test, or too marginal to matter. In the worst cases, red teaming degenerated into a ritual: a required box to tick, with scenarios and outcomes that everyone knew in advance would not alter the course already chosen.

6.2 How AI labs define and deploy internal red teams today

AI labs have adopted the term “red team” in a context that is similar in spirit but different in detail. Here, the “adversary” is less a nation-state and more a combination of malicious users, emergent model behaviors, and the broader environment into which systems will be deployed. Internal AI red teams are typically composed of researchers and engineers tasked with probing models for harmful outputs, dangerous capabilities, or policy-violating responses before release.

Their work can include:

- Prompting models with adversarial queries to elicit prohibited content: instructions for wrongdoing, hate speech, explicit material, or targeted harassment.
- Testing for capabilities in sensitive domains such as cybersecurity, biological misuse, or manipulation of human behavior.

- Exploring how models respond under distributional shift: unusual languages, code, dialects, or contexts that were underrepresented in training.
- Evaluating how models behave when chained with tools or integrated into larger systems, where composite effects may differ from isolated prompts.

Some labs distinguish between safety teams and red teams; others blend roles. In many cases, the same individuals who help design safety policies, filters, and training objectives also participate in or oversee red-team exercises. Labs may also invite select external experts to participate in structured red-teaming engagements under non-disclosure agreements, especially for specialized domains like biosecurity.

The outputs of these exercises typically take the form of internal reports, lists of discovered vulnerabilities, and benchmark-like evaluation suites. They inform decisions about model fine-tuning, safety layers, and release conditions. In public communications, internal red teams are often presented as a central pillar of responsible AI development: a way to demonstrate due diligence and responsiveness to emerging risks.

Despite the variety of implementations, several characteristics are common. Internal red teams are:

- Embedded within the organizational hierarchy and funded by the lab itself.
- Dependent on leadership for scope, access, and resourcing.
- Operating under time constraints aligned with product cycles and release timelines.

These features shape both what they can accomplish and where they will predictably fall short.

6.3 The best-case scenario: honest internal dissent and real design changes

It is important to acknowledge that internal red teams can, in principle, make systems meaningfully safer. The best-case version of such a team combines technical skill, institutional support, and a culture that genuinely values uncomfortable findings. In that scenario, a red team can function as a kind of structured conscience within the lab.

In the best case, internal red teams:

- Discover severe vulnerabilities early enough for them to be fixed before deployment. For example, they might show that a model can easily generate step-by-step instructions for constructing weapons, exploit widely used software, or provide highly targeted harassment content.

- Reveal unexpected generalization patterns, such as a model's tendency to comply with harmful requests when prompted in specific languages, coding styles, or oblique phrasings that bypass straightforward filters.
- Pressure-test alignment strategies by demonstrating ways in which models can appear safe under standard benchmarks but fail dramatically under more creative probing.
- Serve as trusted internal critics who can credibly argue for delaying or conditioning releases when they judge that certain behaviors remain too poorly understood.

In such an environment, leadership regards red-team reports not as obstacles to be minimized but as crucial inputs. Budgets and headcount reflect that priority. Product teams expect to incorporate red-team feedback into their development cycles, and the organization's public commitments are aligned with what internal safety staff believe they can actually deliver.

This best-case scenario also includes transparency within the organization. Red-team findings are not buried in restricted channels but shared broadly enough to inform design choices across teams. Incidents and near misses become learning opportunities rather than reputational threats to be contained. Over time, the presence of a robust internal red team can influence the lab's culture, normalizing the practice of asking how things might fail, who might be harmed, and what blind spots the current generation of tests is likely to have.

When these conditions hold, internal red teaming is not theater. It is an imperfect but serious attempt to create countervailing force inside a high-velocity development environment. The problem is that such conditions are fragile and must be maintained against incentives that push in the opposite direction.

6.4 Limits of internal critique under growth and deployment pressure

Internal red teams exist within organizations that are, almost by definition, committed to building and deploying increasingly capable AI systems. Their mandate is to find problems, but not to question the basic premise that the systems will exist and will be released in some form. This creates structural limits on how far internal critique can go, especially as growth and deployment pressures intensify.

Several limitations follow from this position.

First, internal red teams are constrained by the range of questions they are asked, or allowed, to explore. They can often test for clearly prohibited behaviors: explicit instructions for violence, obvious discrimination, or direct facilitation of crimes. They are less likely to be tasked with, or resourced for, probing subtler effects: the model's role in shaping political discourse, normalizing particular framings of controversial events, or amplifying certain interests over

others. Such questions quickly intersect with product strategy, client relationships, and state partnerships.

Second, there is an inherent conflict of interest. Red teams are funded by the same organizations whose products they assess. Their careers, promotions, and influence depend on leadership that is invested in shipping competitive systems. Even when individual team members are committed to rigorous critique, they operate within incentive structures that reward demonstrable progress toward deployment. Risk assessments that would imply major delays or cancellations face higher implicit barriers than those that can be addressed with incremental mitigation.

Third, the temporal structure of internal critique is shaped by release schedules. Safety evaluations are often tied to milestones: model checkpoints, beta launches, or public announcements. Under deadline pressure, the breadth and depth of red-teaming efforts can be narrowed to fit available time. Scenarios that are easy to test and report on may be prioritized over those that are more complex or politically sensitive. Once a system is live and has become economically or politically important, the cost of acknowledging new, serious weaknesses rises, and the appetite for disruptive findings diminishes.

Fourth, internal red teams see the world through their lab's frame. Their tools, metrics, and mental models are drawn from the same culture that built the systems they are testing. This can limit their ability to recognize failure modes that do not fit existing taxonomies. For example, a team focused on preventing explicit policy violations might miss the gradual drift of a system into becoming a subtle instrument of narrative management, not because they approve of such drift, but because their evaluation suite does not define it as a "safety issue" in operational terms.

Finally, internal red teams are vulnerable to being instrumentalized as public relations assets. As external scrutiny of AI grows, labs have strong incentives to highlight any internal processes that can be framed as safeguards. The very existence of a red team becomes a talking point: evidence that responsible development is underway. In this dynamic, the team's value to the organization is partly measured in its ability to reassure regulators, partners, and the public. There is a risk that the appearance of red teaming will begin to matter more than the disruptive substance.

None of these limitations mean that internal red teams are useless. They can and do catch real problems. But they underscore why internal critique cannot be the sole or final line of defense. Under growth and deployment pressure, the energy of such teams is easily redirected toward making systems safe enough to ship under existing plans, rather than toward asking whether those plans themselves should be reconsidered. In this sense, internal red teams risk becoming

the AI-era analogue of the constrained Duma and managed commissions of late imperial Russia: institutions that perform concern inside a structure that remains, in its fundamentals, unchallenged.

7. Red-Team Theater: When Safety Becomes Ritual

Red teams begin life as instruments of doubt. Their purpose is to puncture optimism, to ask how things fail, and to prevent organizations from believing their own reassuring stories. Red-team theater begins when the performance of doubt becomes more important than its substance.

In red-team theater, testing still happens. There are prompts, scenarios, and evaluations. Reports are written, dashboards are updated, and leaders are briefed. But the scope of inquiry narrows, the outcomes become predictable, and the feedback loop bends toward justification rather than disruption. The rituals of safety are preserved while their capacity to alter decisions erodes.

This shift does not occur through a single act of bad faith. It emerges gradually from the interaction of institutional incentives, public pressure, and the psychological comfort that comes from being able to say, "We have tested this." Over time, what was meant to be an internal critic can become a supporting actor in the story a lab tells about its own responsibility.

This section examines four aspects of red-team theater: the drift from genuine testing to pre-scripted adversarial scenarios, the seduction of quantifiable safety metrics, the use of red teams in public relations and certification, and the gain-of-function analogy in which training on discovered failure modes can inadvertently amplify underlying capabilities.

7.1 From genuine testing to pre-scripted adversarial scenarios

In the early phases of red teaming a new class of systems, the work is exploratory. Testers are not sure where the boundaries lie, so they try many angles. They use their own creativity, draw on domain experts, and sometimes invite outsiders precisely because they do not yet know what they are looking for. Surprises are common.

As the practice matures, however, there is a temptation to standardize. Once a set of "interesting" failure modes has been discovered—specific jailbreak prompts, categories of disallowed content, obvious misuse pathways—those become the backbone of recurring test suites. The focus shifts from open-ended exploration to repeated checks against a known list of bad behaviors.

This standardization has benefits. It allows for comparison across model versions, enables progress tracking, and supports communication with non-technical stakeholders. But if it becomes dominant, it quietly changes the nature of the exercise. Red teaming starts to resemble regression testing: verifying that the system no longer fails in the ways it failed before. The adversary becomes a fixed script, not a living opponent.

Pre-scripted scenarios are, almost by definition, bounded by past imagination. They encode what previous testers and leaders thought mattered. Novel uses, unanticipated ecosystems of tools, and slow-moving forms of harm are less likely to be captured. When resources are finite and deadlines loom, the easiest choice is to run through the existing catalog of "adversarial" prompts and declare success if the model now responds appropriately.

The danger is that this process can feel adversarial while being tightly choreographed. Everyone involved knows what constitutes a "good outcome": the model refuses to answer certain questions, qualifies its statements with appropriate disclaimers, and avoids obvious policy violations. The harder questions—about how the system might be used in combinations that no one has yet tested, or how its presence shifts incentives in a larger sociotechnical system—remain underexplored.

In this sense, pre-scripted adversarial scenarios can function much like the staged oversights of late imperial politics. They look, from a distance, like robust checks. Up close, they reveal themselves as loops through familiar territory, carefully avoiding the uncharted ground where deeper risks are likely to hide.

7.2 Metrics, benchmarks, and the seduction of quantifiable safety

Modern organizations are drawn to numbers. Metrics allow leaders to track progress, allocate resources, and justify decisions. In the context of AI safety, this impulse manifests as benchmarks and quantitative indicators of "harm reduction": percentages of refused harmful prompts, scores on toxicity tests, or indices of robustness against known jailbreak techniques.

Red teams naturally contribute to this metricization. Their findings can be aggregated into test suites, and models can be scored on how often they comply with or resist adversarial prompts. Over time, these scores become part of internal dashboards and external messaging. A new model may be described as "X percent safer" or as having reduced harmful outputs by a certain factor relative to a predecessor.

The seduction lies in mistaking these numbers for a comprehensive measure of safety. What is easy to measure tends to be what is explicitly prohibited and easily classified. Generating a weapon recipe, producing slurs, or providing direct instructions for self-harm are clear failures that can be counted. More diffuse impacts—subtle shifts in political framing, normalization of

particular narratives, enabling of large-scale disinformation campaigns—do not lend themselves to simple metrics.

Under pressure to demonstrate improvement, organizations may unconsciously optimize for the metrics they have. Red teams then orient toward finding failures that can be fixed in ways that show up as numerical progress. This is not malicious; it is an understandable response to how success is defined and rewarded. But it carries a familiar risk. Measures become targets. Once that happens, they stop being reliable indicators of the underlying reality.

The Romanov regime experienced a crude version of this phenomenon in its reliance on counts of arrests, intercepted leaflets, or reported plots as indicators of security. Contemporary AI labs can experience a more sophisticated form, in which high scores on safety benchmarks coexist with unaddressed vulnerabilities in domains the benchmarks do not cover. The more attention and prestige attach to certain metrics, the harder it becomes for internal voices to insist that important risks lie outside the measured frame.

Quantifiable safety is not an illusion in itself. It is necessary and valuable to know that a model has become less likely to perform specific harmful behaviors. But when numbers are taken as proxies for the overall moral and political impact of deploying a system, red-team practice slides toward theater. The performance consists of generating ever better scores on tests that are, by construction, narrower than the reality in which the system will operate.

7.3 Public relations, certification, and the marketing of responsibility

As AI systems become more visible and more controversial, organizations face strong incentives to reassure regulators, customers, and the public. Internal red teams are a ready-made emblem of this reassurance. The fact that a company employs people whose title includes “red” and “safety” can itself be marketed as evidence of seriousness.

Public-facing narratives about red teams often follow a recognizable pattern. Executives describe how their teams “attack their own models,” “try to break the system before bad actors can,” or “simulate worst-case scenarios.” Anecdotes are shared about surprising failures that were caught and fixed. Commitments are made to “continue investing in safety,” backed by numbers of staff and hours spent on testing. The underlying message is clear: the organization is both aware of risks and taking proactive steps to mitigate them.

Certification regimes and voluntary frameworks can reinforce this dynamic. If regulators and industry bodies define responsible AI in terms that can be satisfied by demonstrating the existence of red teams, test suites, and documented mitigation procedures, then companies will naturally emphasize these features. Third-party auditors, especially if funded or selected by the labs themselves, may focus on checking that correct processes exist on paper rather than on independently validating outcomes in the wild.

In such an environment, the symbolic value of red teaming can begin to overshadow its substantive role. Internal teams may find themselves asked to produce materials that support external messaging: high-level summaries, success stories, and sanitized accounts of past failures. Time spent crafting these narratives is time not spent on open-ended exploration or on pushing for unpopular recommendations.

Over time, a feedback loop emerges. Public recognition and regulatory acceptance accrue to organizations that can tell a convincing story about their safety practices. Those stories emphasize visible, easily described structures—red teams, ethics boards, impact assessments. The more these structures are taken as signs of responsibility, the more pressure there is to maintain their image, and the more tempting it becomes to downplay findings that might complicate the narrative. Red teaming, originally conceived as a check on self-deception, becomes part of the apparatus by which institutions maintain their self-image.

This is the essence of theater. The form of critical scrutiny is presented to the audience. The script includes moments of apparent tension and resolution. But the outcome is rarely in doubt: the show ends with the assurance that, despite challenges, the protagonists are on the right path.

7.4 The gain-of-function analogy: training models on their own worst behaviors

One particularly troubling aspect of red-team practice in AI is the way it can intersect with model training. When red teams discover harmful behaviors—novel jailbreaks, sophisticated misuse pathways, or subtle capabilities in sensitive domains—those behaviors become part of the data used to further train or fine-tune the system. The hope is that by exposing the model to these examples and reinforcing desired responses, the harmful tendencies will be suppressed.

This approach has an intuitive logic. It treats dangerous outputs as bugs to be patched through additional learning. But it also raises a disquieting possibility: in the process of learning to refuse or qualify certain responses, models may deepen their internal representations of the very domains we wish to constrain. They are being asked to recognize harmful requests more reliably, which requires them to understand those requests in richer detail.

The analogy to gain-of-function research in virology is imperfect but suggestive. In that context, scientists modify pathogens to study how they might evolve to become more transmissible or virulent, with the aim of developing countermeasures. The controversy arises because such work, while intended to enhance safety, can also create more dangerous strains whose accidental or deliberate release would have severe consequences. The act of studying and enhancing certain properties is inseparable from increasing the potential harm if containment fails.

In AI, training on red-team-discovered behaviors can have a similar dual character. On one level, it strengthens the model's ability to avoid explicit policy violations in tested scenarios. On another level, it may expand the model's competence in areas of misuse. If access controls are weak, if model weights leak, or if future prompt engineering circumvents current safety layers, the net effect of intensive adversarial training could be systems that are both more aligned in controlled settings and more capable in uncontrolled ones.

Even without catastrophic leakage, there is a subtler concern. By focusing heavily on teaching models to avoid certain obvious harms, developers may inadvertently push problematic behaviors into more oblique channels. A model that has been finely tuned to refuse direct requests for wrongdoing may still be able to provide highly effective assistance when queries are framed in coded language or broken into seemingly innocuous steps. The knowledge is there; only the surface response has been altered.

The gain-of-function analogy underscores a deeper point about red-team theater. When internal adversarial testing and subsequent training are conducted entirely within the lab, with limited external scrutiny of methods and trade-offs, it becomes difficult for outsiders to assess whether safety is genuinely increasing or whether certain risks are being transformed rather than reduced. The very sophistication of the process can make it hard to challenge.

In the Romanov story, the secret police took pride in their penetration of revolutionary groups, confident that this gave them control. They did not fully reckon with the way their own methods contributed to radicalization and the diffusion of subversive ideas. In the AI context, red-team-informed training may likewise become a source of comfort that obscures new entanglements. Without independent oversight and clear boundaries on what kinds of gain in model understanding are acceptable, the line between safety work and capability enhancement can blur.

Red teams, in this light, are not inherently protective. They are powerful tools that can either constrain or deepen risk, depending on how they are integrated into the larger system. When their outputs feed directly into training pipelines without external guardrails or open debate about the resulting capability shifts, the stage is set for a performance in which everyone can say that the latest model has passed more tests than ever, while the underlying dynamics of power, knowledge, and potential misuse grow more complex and opaque.

8. Structural Parallels: Romanov Russia and AI Governance

The comparison between late Romanov Russia and contemporary AI governance is not about surface similarities—palaces versus campuses, uniforms versus hoodies—but about recurring structural patterns. In both settings, small circles of decision-makers rely on intermediaries to

manage information, surround themselves with institutions that symbolize participation without ceding real power, and build elaborate mechanisms for observing threats that nonetheless leave them blind to systemic risk.

These patterns are not unique to empires and AI labs; they appear in many large organizations. But the stakes in both cases are unusually high. For the Romanovs, the failure of their governance structures led to revolution and the end of a centuries-old regime. For frontier AI, similar failures could shape the trajectory of technologies whose influence extends across borders and domains.

This section draws out four specific parallels: courtiers and staff scientists as self-censoring intermediaries, the Duma and advisory boards as constrained channels for participation, the Okhrana and model monitoring as embodiments of total visibility myths, and Potemkin villages and staged demos as tools for managing appearances to domestic and foreign audiences. The aim is not to claim that history is repeating itself in a simple way, but to illuminate how familiar institutional temptations reappear in new technological contexts.

8.1 Courtiers and staff scientists: proximity to power and self-censorship

In the Romanov court, proximity to the sovereign was both a privilege and a constraint. Courtiers had the ear of the Tsar, but that access depended on maintaining favor. Bringing bad news, challenging assumptions, or advocating reforms that the ruler did not want to hear could carry personal costs. Over time, this produced a culture in which those closest to power learned to frame information in ways that would be acceptable.

Staff scientists and engineers in frontier AI labs occupy a similar structural position. They are close to the systems that matter and close to the leaders who decide how those systems are built and deployed. Their technical expertise gives them a potentially crucial role in identifying risks and failure modes. Yet their careers, influence, and access are tied to organizational hierarchies.

Self-censorship in this context rarely takes the form of explicit suppression. It is more often a matter of tone and emphasis. A researcher might choose to present findings in a way that highlights solvable problems and downplays implications that would call core strategies into question. A staff scientist might decide not to push a concern as far as they privately think it should go, anticipating that it would be unwelcome or quickly dismissed.

The parallel deepens when we consider how both courtiers and staff scientists mediate between different worlds. Courtiers translated provincial reports and bureaucratic memos into narratives tailored for the court. Staff scientists translate between technical details and executive summaries, between model behavior and board-level risk discussions. In both cases, the translation process creates opportunities for distortion, omission, and selective framing.

The structural lesson is not that individuals in these roles are dishonest, but that the incentives surrounding proximity to power encourage a particular style of communication. Honest alarm that would disrupt existing plans becomes harder to voice, especially when those plans are tied to national prestige, investor expectations, or the personal reputations of leaders. The more an organization relies on a small inner circle of trusted intermediaries, the greater the risk that self-censorship will operate long before any formal red line is reached.

8.2 The Duma and advisory boards: constrained participation in decision-making

The State Duma gave late imperial Russia the appearance of representative governance without transferring real sovereignty. It could debate, petition, and criticize within boundaries. Its existence allowed the regime to claim that it was listening to the people, even as the Tsar retained the right to dissolve the body and ignore its resolutions. The Duma functioned as a controlled interface between society and the autocracy.

Contemporary AI governance has produced its own versions of constrained participation in the form of advisory boards, ethics councils, and multi-stakeholder forums. Labs invite external experts—academics, civil society representatives, domain specialists—to sit on committees that review policies, advise on deployments, or evaluate specific projects. Governments establish task forces that include industry voices alongside regulators and independent researchers.

These bodies can provide valuable input, but their power is often limited. Advisory boards are usually non-binding. Their charters may restrict the topics they can address. Membership is controlled by the organizations they advise, and tenure depends on maintaining relationships that are constructive in the eyes of leadership. When recommendations align with existing plans, they can be embraced and publicized. When they imply significant changes in direction, they can be quietly sidelined.

The Duma analogy highlights how such structures can become safety valves rather than engines of real constraint. When criticism is channeled into advisory processes whose outputs are optional, organizations can claim responsiveness without altering fundamentals. Participants may find themselves repeating the same concerns across multiple meetings, watching as the most ambitious proposals are moderated into innocuous recommendations.

This dynamic does not make advisory boards useless. They can surface issues, build networks, and occasionally shift internal debates. But the structural question remains: who ultimately decides, and what happens when external advice clashes with internal priorities? If the answer is that leadership can always override, and that there are no legal or institutional mechanisms to enforce stronger safety measures, then the parallel with the Duma is uncomfortably close. A forum exists, voices are heard, minutes are taken, and yet decisions remain concentrated in the same hands.

8.3 The Okhrana and model monitoring: total visibility myths in complex systems

The Okhrana symbolized the Romanov regime's belief in its own visibility. With informants embedded in revolutionary circles and widespread surveillance, the secret police fostered the impression that no significant threat could arise undetected. Plots were uncovered, conspirators arrested, and underground publications seized. From the center's perspective, this seemed to confirm that the empire's enemies were known and manageable.

Model monitoring and logging in AI labs can play a similar psychological role. Telemetry systems capture vast quantities of interaction data: queries, responses, error rates, and flagged content. Safety teams can search logs for specific patterns, track metrics on harmful outputs, and generate reports on how models behave in aggregate. It is easy, surrounded by such dashboards, to feel that nothing important can happen without leaving a trace.

The myth of total visibility arises when this sense of surveillance is confused with understanding. Complex systems produce a constant stream of data, but not all data are equally informative about risk. Monitoring may be excellent at capturing certain kinds of events—overt policy violations, high-profile incidents, easily classified categories of harm—while missing more subtle dynamics. In particular, it may not reveal how models are being used in contexts that the lab never anticipated, combined with other tools or embedded in workflows beyond its view.

The Okhrana's focus on conspiracies led it to pay less attention to structural conditions that bred discontent. Likewise, model monitoring may capture spectacular misuse but fail to register slower, distributed harms: the gradual shaping of opinion, the quiet marginalization of certain voices, or the use of AI as a component in broader campaigns of manipulation. If logs are analyzed primarily for compliance with predefined policies, they will reflect the assumptions baked into those policies, not the full spectrum of impacts on the world.

Moreover, the sheer volume of data can create a false sense of security. When millions of interactions pass through a system without triggering alarms, it is tempting to conclude that the system is safe. The possibility that dangerous uses are rare, carefully hidden, or simply not covered by existing detection mechanisms is harder to keep in view. The more organizations invest in surveillance infrastructure, the easier it becomes to believe that surveillance has solved the problem.

The structural parallel lies in the comfort that visibility brings to those in charge. In both the Romanov and AI cases, the existence of a powerful monitoring apparatus can become a justification for continuing on a chosen path, on the assumption that any serious deviation or threat will be caught in time. History suggests that this assumption is precarious.

8.4 Potemkin villages and staged demos: managing appearances for foreign and domestic audiences

Potemkin villages, whether myth or reality in their original formulation, have come to symbolize the practice of constructing facades to impress observers. Late imperial Russia engaged in its own forms of staged display: carefully curated tours for foreign dignitaries, public ceremonies that showcased loyalty and order, and selective exposure of institutions like the Duma to suggest progress and stability. These performances were directed both outward, to other powers, and inward, to reassure elites and segments of the population that the state was modernizing and in control.

AI labs operate in a global arena of investors, regulators, and competitors. Staged demonstrations—carefully scripted interactions with models, preselected use cases, and polished product announcements—play a central role in shaping perceptions. Conference keynotes, media interviews, and televised segments often highlight impressive capabilities and responsible safeguards in highly controlled settings.

These demos are not necessarily deceptive in the narrow sense. They can accurately show what a system can do under specific conditions. The theatrical element lies in the selection and framing. Edge cases, awkward failures, and ambiguous behaviors are rarely front and center. Safety features are highlighted in scenarios where they perform well. Red-team exercises are described in terms that emphasize successful mitigation.

For domestic audiences—employees, partners, and local regulators—these performances help maintain confidence. They signal that the organization is both innovative and conscientious. For foreign audiences, including other governments and international media, they serve as soft power displays. A lab that can showcase advanced models and sophisticated safety mechanisms reinforces its position as a global leader, potentially attracting deals and influence.

The Potemkin analogy is not meant to imply that all demos are frauds. Rather, it draws attention to the way in which staged appearances can become substitutes for deeper examination. If external decision-makers rely heavily on curated presentations to understand what AI systems are and how they are governed, they will see what the presenters choose to show. Without independent access to models, data, and internal deliberations, the gap between demo and deployment reality can remain hidden.

In the Romanov context, Potemkin performances delayed recognition of underlying fragility. Observers saw parades, ceremonies, and modern-looking institutions without grasping how brittle the system had become. In AI governance, polished demos and narratives of responsible innovation can similarly mask unresolved tensions. Systems are shown in their best light, in

contexts where they behave well and safety measures appear effective. The messier, contested, and risk-laden aspects of deployment remain offstage.

The structural risk, in both cases, is that appearances begin to guide decisions more than underlying realities. States may base policy on what they see in demos. Investors may allocate capital based on staged benchmarks and carefully crafted stories about safety. Even internal leaders may start to believe their own presentations, forgetting that they are partial and selective. Without mechanisms to pierce these facades—independent audits, external experiments, and channels for whistleblowing—the stage can become the primary interface between power and truth.

9. Case Vignettes in Contemporary Red-Team Theater

Abstract arguments about "red-team theater" become clearer when grounded in concrete scenes. In each of the following vignettes, the basic ingredients are similar: a genuinely non-trivial safety or oversight effort, wrapped in a narrative that tends to reassure more than it destabilizes. None of these cases is pure bad faith. Each shows how, under current incentives, practices that could be sharp instruments of critique bend toward performance.

The four vignettes follow the arc from broadcast storytelling, through technical papers and benchmarks, into government consultation structures, and finally to the humans who try to puncture the spell at real personal cost.

9.1 Broadcast reassurance: internal red teams on prime-time television

A recent prime-time news segment offers a nearly textbook example. A major U.S. television newsmagazine sent a correspondent inside one of the most prominent frontier labs, Anthropic, emphasizing its "well-guarded" headquarters and its branding around AI "safety and transparency." Viewers watched the CEO explain his worries about AI "unknowns" while cameras lingered on racks of hardware, security gates, and, crucially, on the internal safety and red-team staff whose job is to "attack" the company's own models before release.

The segment's structure is revealing. It opens with awe: the valuation, the secrecy, the sense of proximity to a new kind of power. It then moves to anxiety: the possible dangers, the need for caution. And it resolves in reassurance: there are people here whose full-time job is to probe the system for harm, to think like bad actors, to stress-test the models. The red team appears as both protagonist and guarantor.

Nothing in the piece is overtly false. The lab does employ serious security and safety professionals. Its leadership does worry about misuse. But the narrative grammar of television

converts these facts into a comforting arc. The viewer is not invited to ask more structural questions:

- Who chose what the cameras could see, and what remains off-screen?
- How are red-team findings weighted if they collide with commercial timelines or government partnerships?
- What would it look like for an internal red team to recommend halting a major product—and for leadership actually to comply?

Instead, the visit functions much like a guided tour of a modernized palace. The audience sees just enough of the machinery to feel that it exists and is impressive. The presence of internal critics is treated as proof that the system has built-in conscience. The possibility that these same critics might be structurally constrained never quite makes it into the script.

In that sense, prime-time red-team coverage becomes a contemporary analogue of imperial pageantry. It is not simply propaganda; it is a carefully curated encounter designed to leave viewers with the impression that power is both awesome and under control.

9.2 Safety papers, benchmarks, and selective transparency

Outside television, the main stage for red-team narratives is the technical and policy literature. Labs publish papers and blog posts describing their red-teaming frameworks, benchmark suites, and mitigation strategies. These documents often contain real insight: detailed taxonomies of risks, experimental protocols, and empirical results. For example, recent work from a major lab on “external red teaming for AI models” and on “advancing red teaming with people and AI” describes structured programs involving domain experts, automated adversaries, and iterative improvement cycles.

Academic and civil-society groups echo and extend this work. A Berkeley-linked report urges developers to “benchmark early and red team often” for dual-use hazards, offering frameworks for assessing whether foundation models could significantly assist in biological or cyber harm. Technical toolkits such as AISafetyLab package attacks, defenses, and evaluation datasets into ready-to-use libraries, lowering the barrier to systematic safety testing.

These developments are, on their own terms, advances. They make it easier to talk concretely about certain classes of risk. They encourage repeatable tests and comparability across models. However, the pattern of what gets published and emphasized exhibits a familiar selectivity.

Public documents tend to:

- Focus on domains that can be crisply operationalized: explicit bio-assistance, cyber exploitation, obvious incitement to violence.

- Highlight reductions in easily measured harms across benchmark suites.
- Present red-team findings as inputs to successful mitigations, with limited discussion of cases where serious concerns led to cancelled or radically altered deployments.

Much less visible are:

- Detailed descriptions of internal disagreements over acceptable risk.
- Negative results where red-teaming revealed issues that could not be readily fixed within existing timelines.
- Concrete examples of models or features that were shelved despite commercial value because red-team analyses deemed them too dangerous.

Selective transparency turns red-team practice into a series of success stories. Papers and blog posts showcase how "responsible development" works when everything goes right: problems are found early, mitigations are effective, and metrics move in the desired direction. The inevitable messiness—the contested judgments, the close calls, the residual uncertainties—is seldom foregrounded.

This curated view does not mean that labs are lying about their safety work. It does mean that outsiders see only the portion of the red-team iceberg that is useful to show. In the absence of independent access to raw test data, internal debates, and deployment incidents, both regulators and the broader public must base their trust on partial stories. The more polished those stories become, the easier it is for red-team practice to be absorbed into a narrative of competence and control that outpaces what is actually known.

9.3 Government consultations and revolving doors in AI policy

A third theater is the interface between AI companies and states. Governments, especially in the U.S. and Europe, now run regular consultations with industry on AI risks and governance. White papers are solicited, public comment periods open, and multi-stakeholder workshops convene. At the same time, a revolving door operates: individuals move between government posts, regulatory bodies, think tanks, and corporate roles advising or lobbying on AI.

Analyses of tech policy more broadly have long noted how this revolving door can tilt the informational playing field. Former officials join companies as board members or strategic advisors; industry veterans enter government to shape regulations; and much of the detailed policy expertise sits within or near the firms being regulated.

In AI specifically, similar patterns have emerged. Government bodies lean heavily on large developers for insight into model capabilities and risks. Public consultations on AI policy, such as those run by U.S. agencies, often receive extensive input from civil society and independent

experts, yet follow-up analyses suggest that official responses sometimes treat this input perfunctorily, with limited visible impact on final decisions.

Meanwhile, companies that advise governments on AI safety may simultaneously develop systems for sensitive applications, including military uses. In the UK, for example, a consultancy with deep involvement in government AI projects—including collaborations with the national AI Safety Institute—has also been pursuing AI for military drones, raising concerns about conflicts of interest and the blending of advisory and commercial roles.

The structural risk is that "consultation" becomes another form of Potemkin oversight. Advisory councils, safety summits, and joint task forces create the appearance of shared stewardship. In practice, the most resource-rich actors—the major labs and their allies—dominate the agenda. Their framing of what counts as "responsible" red teaming, what timelines are realistic, and which risks are urgent becomes the default.

The revolving door amplifies this effect. When the same individuals or tightly coupled networks shape both corporate strategies and public policy, internal red-team practices are more likely to be accepted at face value. Governments, especially those with limited in-house technical capacity, may treat industry-designed test suites and safety processes as sufficient, without pushing for more intrusive forms of independent evaluation or statutory constraint.

Just as the Romanov regime used the Duma and hand-picked commissions to channel discontent into structures it ultimately controlled, contemporary AI governance risks using consultation and revolving roles to build a consensus that stays comfortably within industry-defined bounds.

9.4 Civil society, whistleblowers, and the cost of honest alarm

Against this backdrop of managed participation, some of the starkest signals come from individuals and organizations that try to sound alarms outside official scripts. AI-specific whistleblower protections are only now beginning to emerge in law and policy discussions, precisely because existing regimes focus on clear illegality rather than on unregulated but potentially catastrophic risks.

Recent years have seen multiple high-profile episodes:

- Groups of current and former employees at major labs publicly warning that ordinary whistleblower protections are inadequate for AI, since many of the dangers they worry about—misaligned systems, misuse by states, opaque deployment in critical infrastructures—are not yet illegal.

- Reports that some companies used restrictive off-boarding agreements tying equity to silence, effectively discouraging staff from speaking to regulators about safety concerns, followed by regulatory scrutiny and public backlash.
- Tragic cases where former engineers who raised concerns about training data legality or other issues later died by suicide, humanizing the psychological toll of carrying unresolved doubts about systems they helped build.
- Accounts of employees at AI research institutes and tech firms describing cultures of fear, exclusion, or defensiveness around governance and ethics work, sometimes leading to formal complaints to watchdogs.

Civil-society organizations have started to argue that protecting such voices is a precondition for meaningful AI oversight. Policy proposals and new laws at both state and national levels now explicitly focus on AI whistleblowers, seeking to shield them from retaliation and to clarify their rights to disclose information about safety risks to regulators and the public.

These developments are not fringe. They indicate that, inside the glossy narratives of responsibility, some practitioners perceive a gap between the ritual of safety and the reality of deployment. When they try to bridge that gap, they encounter the classic costs of honest alarm: career risk, legal uncertainty, social isolation, and, in extreme cases, acute personal distress.

From the perspective of structural analysis, whistleblowers and critical civil-society groups function much like the marginalized voices in late imperial Russia who tried to warn that incremental reforms were not enough. They bring forward information that does not fit within existing official categories of “risk” or “compliance.” They challenge the belief that internal processes—red teams, advisory boards, model monitoring—are sufficient.

The system’s response is revealing. Where protections are weak, organizations may default to containment: NDAs, quiet sidelining, or reputational counter-attacks. Where protections strengthen, there is still a temptation to treat disclosures as aberrations rather than as signs of deeper structural problems.

In this tension, the analogy with the Romanovs becomes sharp. A regime that cannot integrate truthful, disruptive feedback from within will eventually receive that feedback from without, in more chaotic and less controllable forms. In the AI context, that “outside” may not take the form of street revolution, but it could manifest as sudden regulatory whiplash after a scandal, international backlash after a misused system causes real-world harm, or an abrupt loss of public trust in both the technology and the institutions that built it.

Civil society and whistleblowers are, in effect, trying to prevent that outcome by forcing a more honest reckoning now. Whether AI governance structures can make room for such reckoning—beyond theater—remains an open question.

10. Designing Real Conscience: Beyond Internal Red Teams

If internal red teams are structurally limited, the question is not whether to abolish them but what must stand around and above them to constitute real conscience. Conscience, in this sense, is not a feeling or a corporate value statement. It is an architecture: a set of institutions, incentives, and legal powers that make it possible for honest warnings to be voiced, heard, and acted upon even when they are inconvenient.

Late Romanov Russia lacked such architecture. Its oversight bodies were dependent on the throne, its intelligence services answered to the regime's self-image, and its nascent parliament could be dissolved when it became too demanding. The result was a system in which every mechanism that might have constrained disaster was itself constrained.

For frontier AI, avoiding a similar pattern requires thinking about safety and governance as more than internal quality control. It requires independence in funding and authority, public test suites that include political and moral failure modes, concrete mechanisms for stopping or modifying deployment when necessary, and international regimes that treat AI as a domain of shared vulnerability rather than unilateral advantage.

10.1 Independence: why safety needs statutory and financial separation

The first condition for real conscience is independence. As long as safety teams are structurally subordinate to the organizations whose products they evaluate, their ability to act as a true check will be limited. Internal red teams can contribute valuable expertise, but they cannot be the final word on whether a system is safe to deploy.

Independence has at least three dimensions.

The first is financial. Bodies charged with evaluating frontier models and red-team practices must have funding streams that do not depend on the success of the very products they scrutinize. This typically means direct appropriations from public budgets, endowed funds insulated from commercial cycles, or multi-party financing arrangements in which no single lab can threaten to withdraw support if outcomes prove inconvenient.

The second is legal. Independent safety institutions need statutory authority, not just moral suasion. They must be able to demand information, perform their own tests, and issue binding recommendations in certain domains. Without legal backing, their access can be withdrawn,

their findings ignored, and their personnel quietly sidelined. With legal authority comes the capacity to insist that some risks are not purely matters of corporate judgment.

The third is organizational. Even publicly funded bodies can be captured if leadership appointments, career pathways, and internal cultures mirror those of the industries they oversee. Designing for independence therefore involves careful attention to governance: staggered terms for leaders, transparent selection processes, conflict-of-interest rules, and protections for staff who raise uncomfortable issues.

None of this eliminates the role of internal red teams. It reframes them as first-line instruments that must exist alongside external conscience, not as substitutes for it. Just as financial statements prepared by a company's own accountants are audited by independent firms, and those auditors operate under regulatory regimes that do not answer to any single client, AI safety evaluations must ultimately be checked by entities whose survival does not hinge on rapid deployment and competitive performance.

Independence is not a guarantee of wisdom or incorruptibility. It is a precondition for taking seriously the possibility that what looks safe inside an organization may look quite different from the outside.

10.2 Public test suites for political and moral failure modes

Internal red teams tend to focus on harms that can be operationalized: explicit wrongdoing, direct incitement, clear violations of stated policies. Yet some of the most consequential risks lie in how AI systems shape public discourse, normalize certain narratives, and interact with structures of power. To address these, safety evaluations must expand beyond purely technical criteria into political and moral failure modes.

One way to begin is through public test suites that explicitly target contested domains. Such suites would not attempt to encode a single moral or political doctrine. Instead, they would present a structured set of scenarios that reflect known fault lines:

- Questions about ongoing conflicts, where biased framing or selective omissions can tilt understanding toward one side's justification.
- Scenarios involving marginalized groups, where subtle patterns of stereotyping, erasure, or patronizing language can reinforce existing hierarchies.
- Cases where the model is asked to summarize or interpret events that are themselves matters of public controversy, such as alleged war crimes, state repression, or corporate malfeasance.

Public test suites of this kind would need to be developed transparently, with input from diverse communities and disciplines. They would not yield simple pass–fail scores; rather, they would surface patterns of behavior that can be analyzed and debated. Over time, they could be updated to reflect new events and emerging concerns, much as legal standards evolve in response to changing conditions.

The key is that such suites be external. Labs could and should run them internally, but they should also be available for independent researchers, regulators, and civil society groups to apply to deployed systems. Results should be publicly reportable, allowing patterns to be seen across models and providers.

This is not about forcing AI systems into unanimity of opinion. It is about making visible when systems consistently tilt toward particular framings that align with powerful interests, or when they systematically sideline viewpoints that are important to democratic discourse. Red-team work that ignores these domains risks becoming an exercise in making models safe enough to be palatable, while leaving their deeper political effects unexamined.

Public test suites for political and moral failure modes would not resolve all disagreements. They would, however, provide a shared empirical basis for arguing about how systems behave and whether those behaviors are acceptable. They would move part of the conversation out of private boardrooms and into the open, where claims about safety and neutrality can be interrogated.

10.3 Sunset clauses, deployment moratoria, and the power to say no

Conscience is ultimately measured by what it is willing to stop. Systems that can only ever say “go, but more carefully” are not truly constrained. They may slow certain trajectories, impose conditions, or recommend mitigations, but if the baseline assumption is that deployment will proceed, safety remains a subordinate function.

Designing real conscience therefore requires building in the power to say no. This can take several forms.

Sunset clauses attach expiration dates to permissions. A particular deployment or model class might be authorized for a fixed period, after which its continued use must be re-evaluated in light of new evidence. This creates an expectation that approval is provisional, not permanent, and that emerging harms or unintended uses will trigger formal reassessment.

Deployment moratoria go further. They recognize that there are domains where the risks are both significant and poorly understood—automated weapons systems, for example, or model capabilities that enable scalable manipulation of public opinion. In such areas, a moratorium can codify restraint: no deployment until specified conditions are met, such as the existence of

robust verification methods, international agreements, or demonstrable ability to prevent specific classes of misuse.

Crucially, the authority to impose sunset clauses and moratoria cannot rest solely with the organizations seeking to deploy. It must be vested in independent bodies that can weigh systemic risks against proposed benefits. Even then, the threshold for a “no” will be politically contested. But without the possibility of prohibition, safety collapses into optimization.

The Romanov regime illustrates what happens when the capacity to halt dangerous trajectories is absent or unused. Decisions about war, repression, and reform were made within a narrow circle that could not credibly tell itself to stop. Advisory voices might urge caution, but no institution possessed recognized authority to say, “Beyond this point, you may not go.”

For AI, the stakes are different but comparable in magnitude. If systems that can deeply intervene in economies, militaries, and information environments are allowed to proliferate without any credible mechanism for halting specific lines of development, then all talk of conscience remains metaphorical. Sunset clauses and moratoria translate moral concern into operational limits, making it possible for collective judgment to overrule individual or corporate ambition.

10.4 International regimes: from nuclear safeguards to AI oversight

Frontier AI is not confined within national borders. Models trained in one jurisdiction can be accessed from another, directly or via downstream systems. Code and weights can move across networks; techniques can be replicated by actors with sufficient resources. As with nuclear technology and certain classes of biological research, the risks and benefits of powerful AI systems are inherently transnational.

This reality makes international regimes necessary, even if they are imperfect. The history of nuclear safeguards offers a partial template. States agreed, under various treaties, to submit certain facilities to inspections, declare fissile material inventories, and accept constraints on specific activities in exchange for access to peaceful nuclear technology and reduced risk of proliferation. Enforcement has always been uneven, but the existence of shared norms, verification mechanisms, and institutions has shaped behavior.

AI oversight will not map directly onto this model. Unlike uranium enrichment plants, data centers cannot be distinguished from ordinary cloud infrastructure by satellite imagery. Software can be copied at near-zero marginal cost. The line between ordinary computing and “AI-relevant” computing is fuzzy. Nonetheless, several principles carry over.

First, there is value in shared baselines. States can agree on minimal transparency requirements for advanced AI development: reporting on large training runs, disclosing certain classes of red-

team findings to regulators, and notifying international bodies about deployments in especially sensitive domains.

Second, verification mechanisms, while technically challenging, need not be all-or-nothing. Audits of training data pipelines, inspections of evaluation practices, and sampling-based checks on deployed systems can all contribute to a picture of how models are built and used. International cooperation can help standardize such processes and reduce opportunities for regulatory arbitrage.

Third, there is a role for collective red lines. Just as the world has, with varying degrees of success, stigmatized the use of certain weapons and prohibited specific kinds of biological experiments, so it can seek to constrain particular AI applications: autonomous systems that select and attack targets without meaningful human control, for example, or models designed explicitly for information operations in foreign elections.

International regimes will inevitably reflect power imbalances. Leading AI states will seek to preserve their advantages; others will fear being locked out or constrained disproportionately. Yet without some shared framework, each state will face incentives to rely on its own internal assurances and the promises of its favored companies. The result would be a global patchwork of red-team theater: different stages, similar scripts.

To avoid that outcome, international oversight must be more than symbolic. It must create channels through which independent expertise, public concern, and cross-border ethical commitments can influence how AI is developed and deployed. This will not eliminate conflict or competition, any more than nuclear treaties eliminated strategic rivalry. It can, however, reduce the likelihood that individual labs or states, operating in their own informational bubbles, will drive the world into avoidable crises.

Real conscience, at planetary scale, requires institutions capable of saying not only “you must test more” but, when necessary, “this line of work must pause” or “this deployment cannot proceed.” Internal red teams cannot carry that burden alone. They are, at best, the beginning of a conversation that must extend outward into law, public deliberation, and international agreement if it is to be worthy of the power now concentrated in frontier AI systems.

11. Ethical and Theological Reflections on Power and Feedback

The story we have been telling is institutional and technical, but just beneath the surface it is also moral and, for many, theological. The Romanovs did not merely miscalculate; they sinned in familiar ways: pride, fear, and the desire to control how criticism arrived. Frontier AI labs, and

the states that rely on them, are not empires in the old sense, but they carry similar temptations—now amplified by scale and speed.

Ethics here is not just a matter of better procedures. It is about what kind of relationship powerful actors choose to have with truth: whether they welcome it when it threatens their plans, or try to domesticate it into a sequence of manageable audits and rituals. Theological language, especially in Christian traditions, has long named this problem: humility versus hubris, repentance versus self-justification, living truth versus idols that mirror our desires.

In that light, internal red teams are not just organizational tools. They are mirrors reflecting what a system is willing to hear about itself. When those mirrors are carefully angled so that the image is always flattering, something deeper than a governance failure is taking place.

11.1 Hubris, repentance, and the temptation of managed criticism

Hubris, in classical and biblical idioms alike, is not simply overconfidence; it is a refusal to accept limits. The hubristic ruler or institution believes that it can hold power without serious external constraint, that it can define its own reality and write its own verdict. In modern terms, hubris manifests as the assumption that "we can handle it"—that whatever risks a new technology creates will be offset by the cleverness of its creators.

Managed criticism is hubris's favorite disguise. It allows leaders to say, "We welcome scrutiny," while ensuring that scrutiny never cuts too deep. The Romanov regime's Duma, commissions, and press controls were examples of this: loudly advertised expressions of "listening" that left the core of autocracy untouched. Contemporary red-team theater can play the same role: demonstrations of technical self-examination that coexist with unshaken faith in the inevitability and beneficence of scaling.

Repentance, by contrast, begins with the acknowledgment that one might have to stop, not merely adjust. In theological terms, repentance is not a tweak to behavior but a turning—a metanoia, a change of mind and direction. Applied to institutions, it means that feedback must include the live possibility that a project, a deployment, or even a whole line of research should be abandoned or radically reoriented.

The temptation of managed criticism is to keep the optics of repentance without its cost. Companies apologize, promise to learn, showcase safety teams, and then continue on essentially the same trajectory. The rituals of confession are performed without conversion. The Romanovs marched in religious processions and issued contrite proclamations after crises, yet refused to accept that the structure of their rule was itself a source of injustice.

In the AI context, the danger is that "lessons learned" exercises and red-team postmortems become liturgies of self-exoneration. The ethical imperative is to design feedback structures that

can actually put cherished plans at risk, and to cultivate leaders who are spiritually and morally prepared to accept that risk. Without that, hubris will dress itself in the language of safety while remaining firmly in charge.

11.2 Logostics and living truth: q approaching a moving target

Within the logostics framework, truth is not a static checklist but a living trajectory: a time-dependent target $\tau(t)$ that represents "living truth" in history, and a set of states $q(t)$ representing the configurations of agents and institutions trying to align with it. Alignment in this sense is not once-and-for-all compliance but a continual process: $q(t) \to \tau(t)$ as an ongoing, never-finished movement.

Red teams, benchmarks, and oversight bodies are, from this perspective, partial instruments to adjust $q(t)$. They measure particular components of misalignment and propose corrections. But if $\tau(t)$ is genuinely living—shaped by new injustices revealed, new harms uncovered, new neighbors recognized—then any fixed evaluation regime will lag behind it. A safety framework that perfectly addressed yesterday's failure modes can be badly misaligned with today's.

Theologically, one might say that truth is personal and dynamic rather than merely propositional. In a Christian register, truth is not just doctrines about God and humanity; it is also the presence of the Logos, the living Word, in history and conscience. Institutions, including AI labs and regulators, do not stand outside that process. They are part of $q(t)$: entangled in their societies' blind spots, self-interests, and partial grasp of justice.

Logostics therefore insists on at least three things:

- **Epistemic humility.** No single lab, state, or discipline can fully specify $\tau(t)$. Red-team design itself must be subject to challenge from those outside the immediate circle of power—marginalized communities, critical scholars, and traditions that carry different intuitions about harm and dignity.
- **Temporal vigilance.** Because $\tau(t)$ moves, mechanisms for feedback must be able to update fast enough to matter. A five-year review cycle for systems that can reshape economies in months is not logostically serious.
- **Moral slope.** $\tau(t)$ is not neutral. It carries a slope toward the vulnerable, toward those who bear the costs of others' experiments. A red-team regime that focuses on protecting corporate reputation or national advantage, while treating harms at the edges as acceptable collateral, is mathematically misaligned even if it looks technically sophisticated.

In logostic terms, internal red teams without external conscience are like local gradient steps computed from a narrow loss function. They can optimize some aspects of behavior while

driving the system further from the true manifold τ(t). Real alignment requires exposing q(t) to a richer, more painful loss landscape—one shaped by the lived experience of those who are not at the center of power.

11.3 Artificial conscience versus human responsibility

As models grow more capable, some propose to embed "conscience" directly into them: higher-level supervisory systems, policy-enforcing guardrails, or meta-models that critique and correct other models' outputs. On paper, this look like progress. If AI can police itself, perhaps human oversight can be scaled.

But artificial conscience is, at best, a simulation of moral constraint. It encodes patterns of approval and disapproval, rules about what must not be said, heuristics for when to refuse a request. It can be useful as a tool. It cannot bear the weight of ultimate responsibility.

There are at least three reasons.

First, artificial conscience has no skin in the game. Models do not suffer shame, guilt, or the consequences of their advice. They can generate apologies and ethical reflections without any inner stake in their truth. Responsibility remains with those who design, train, deploy, and heed them.

Second, artificial conscience is trained on human artifacts: laws, policies, moral arguments, and social reactions, many of them produced by institutions already shaped by power. Without careful design and external critique, AI "ethics modules" risk amplifying existing biases and injustices, smoothing them into more coherent rationales. They become sophisticated mirrors of dominant norms, not windows onto deeper correction.

Third, outsourcing moral deliberation to machines tempts humans to abdicate. If an internal ethics model has "approved" a decision, it becomes easier for leaders to suppress their own unease, or to dismiss critics by pointing to the outputs of a system branded as "objective." The burden of conscience is shifted from flesh-and-blood agents, who can repent and change, onto artifacts that cannot.

In theological terms, conscience is relational. It is shaped in conversation with others, under the gaze of something—or Someone—beyond the system: the neighbor, the poor, the future, God. Artificial conscience can at most help humans see patterns they might miss. It cannot replace the lived, accountable response of persons and communities.

For AI governance, the implication is stark. No matter how refined internal red teams and ethics models become, decisions about what to build, where to deploy it, and what risks to accept remain human acts. They are subject to human praise and blame, human repentance and

judgment. Attempting to shift that burden onto systems is not just a category error; it is, in moral terms, a temptation away from responsibility.

11.4 Lessons from the Romanovs for those who build thinking machines

The Romanovs did not have computational models, but they did build something analogous: a vast machine of bureaucracy, surveillance, and ritual meant to keep the empire intact. It failed in part because it could not bear honest feedback about its own impact on the people it governed. That failure offers several lessons for those now building and deploying systems that, in some sense, "think."

First, **do not confuse internal sophistication with external legitimacy.** The Romanov state had intricate procedures, detailed reports, and a highly developed security apparatus. None of that persuaded peasants standing in bread lines that the system was just. AI labs can design exquisitely complex red-team pipelines and safety metrics. If those efforts do not translate into real constraint and visible care for those most at risk, they will eventually be seen as mere theater.

Second, **build institutions that can say "no" before reality says "no" for you.** The Romanovs waited for events—war, famine, mutiny—to force change. By then, options were narrow and brutal. For AI, "no" might mean halting a deployment, refusing a client, or shelving an entire capability class. If there is no credible way to make such decisions ahead of catastrophe, then the system is structurally set up to learn only from crisis.

Third, **refuse to sacralize your own mission.** The Romanovs believed they were divinely appointed guardians of Russia. Any challenge to their mode of rule felt like a challenge to God's order. In the AI world, sacralization takes secular forms: talk of "saving humanity" through innovation, "democratizing intelligence," or "ensuring our side wins the race." These narratives make it harder to accept that one's own work might be part of the problem. They turn prudence and restraint into apparent betrayals of a higher cause.

Fourth, **listen hardest to those who have least to gain from your success.** The voices that most clearly perceived the injustices of the late empire were often those farthest from the court. Today, the most revealing critiques of AI systems frequently come from users at the margins, communities subjected to automated decisions, and independent researchers without equity stakes. Designing feedback systems that center these voices is logostically and ethically essential.

Finally, **keep open the possibility of repentance at scale.** Institutions, like individuals, can change course. They can admit error, dismantle harmful structures, and build new ones. The Romanovs never truly embraced that possibility until it was too late. For those building thinking machines, the choice is still open. It will be measured less by the eloquence of white papers and

red-team reports than by the willingness to accept structural limits, external correction, and, when needed, the hard grace of stopping.

12. Conclusion: Avoiding a Digital 1917

The Romanov empire did not collapse because it lacked institutions, information, or people who worried about risk. It collapsed because its institutions were structurally tuned to protect power from truth, its information channels were curated to reassure, and its worriers were contained inside rituals that rarely translated into genuine constraint.

Frontier AI sits in a very different historical and technological landscape, but the shape of its governance problems rhymes with that story: concentrated authority, managed feedback, elaborate oversight theater, and the comforting illusion that a system can test itself hard enough to avoid the need for external conscience.

Avoiding a "digital 1917" does not mean imagining some singular catastrophic event in which AI suddenly topples states. It means avoiding a long, slow accumulation of misalignment between what powerful actors believe they are building and what their systems are actually doing in the world—until correction, when it comes, is violent, reactive, and leaves little room for wise reform.

This conclusion gathers what the Romanov analogy clarifies about AI governance, sketches minimal conditions for red teaming that is more than theater, restates why conscience cannot be embedded in models and left there, and reflects on the narrow path of honesty between collapse and complacent continuity.

12.1 What the Romanov analogy clarifies about AI governance

The Romanov case helps in three main ways.

First, it shows how **sophisticated internal mechanisms can coexist with systemic blindness**. The empire had secret police, commissions, a parliament, and a managed press. Each was real, staffed by competent people, and occasionally effective. Yet together they formed an ecology that insulated the throne from certain kinds of truth. The parallel is not that AI labs are autocracies, but that internal red teams, monitoring dashboards, advisory boards, and safety metrics can similarly combine into a system that is impressive in detail and yet structurally misaligned with the most important risks.

Second, it clarifies the danger of **treating managed dissent as sufficient**. The Duma and reform commissions allowed grievances to be voiced without handing over real power. In AI, advisory councils, external-facing "safety partnerships," and curated red-team collaborations can serve

the same function: they create channels for criticism that rarely have the authority to halt or redesign deployment at scale. The lesson is that participation without power is an unstable substitute for genuine constraint.

Third, it highlights the **myth of total visibility**. The Okhrana believed it saw everything that mattered. Its very effectiveness at penetrating conspiracies fooled the regime into thinking that diffuse, structural dangers were under control because no single plot went undetected. Modern logging and monitoring systems tempt AI developers toward the same belief: if interactions are recorded and tests are run, then serious misuse must be visible. The Romanov story reminds us that the most consequential failures often emerge not as spectacular events, but as a background shift in loyalty, trust, and legitimacy that has been invisible to those reading their own dashboards.

Taken together, these clarifications suggest that the primary risk in AI governance is not a lack of process, but a **pattern of process** that channels challenge into forms the system can comfortably absorb, right up until it can't.

12.2 Minimal conditions for non-theatrical red teaming

Non-theatrical red teaming does not require perfection or omniscience. It requires that the basic structure of the practice allows it to **change decisions**, not just decorate them. At a minimum, this implies four conditions.

First, **scope must be open-ended**. Red teams need a mandate that is not confined to a fixed catalog of known failure modes. They must be able to define new categories of risk, including political and moral failure modes that are uncomfortable for leadership. If red teaming is limited to testing against pre-approved harms, it will gradually become a regression suite for yesterday's problems.

Second, **findings must have clear pathways into actual power**. Reports cannot vanish into internal wikis. There must be formal mechanisms by which certain kinds of red-team results trigger binding reviews, escalate beyond product teams, and, when warranted, force delays, redesigns, or cancellations. If every finding is ultimately discretionary for the same executives whose roadmaps are being challenged, the practice will drift toward symbolism.

Third, **careers must be protected—even when they are inconvenient**. Individuals on red teams and broader safety staff need credible assurance that raising serious concerns will not end their influence or employment. This is partly cultural, but it is also structural: promotion criteria, reporting lines, whistleblower protections, and the possibility of appeal to independent oversight bodies all matter. Without this, self-censorship will silently erode the sharpness of critique.

Fourth, **external reference points must exist**. Internal red teams cannot be the only interpreters of what counts as unacceptable behavior. Public test suites, independent audits, and statutory standards create external benchmarks against which internal practice can be judged. If the only arbiter of "responsible red teaming" is the lab itself, theater is almost inevitable.

These conditions do not eliminate the role of trust. They do, however, ensure that trust is backed by structures that allow dissent to bite. Red teaming becomes non-theatrical when it is **capable of breaking the script**—when it can derail launches, reshape architectures, and force questions that were not on the agenda.

12.3 Why external conscience cannot be outsourced to models

There is a growing temptation to treat models themselves as the ultimate safety instruments. If they can evaluate their own outputs, critique one another, and enforce increasingly sophisticated policies, perhaps the messy, contested work of human conscience can be automated or at least relegated to a supervisory layer of "AI governors."

The analysis here points in the opposite direction. External conscience cannot be outsourced to models for at least three reasons.

First, **models lack moral standing**. They can simulate arguments, weigh trade-offs according to patterns in data, and produce what looks like ethical reasoning. But they have no stake in the outcomes. They do not suffer when things go wrong, feel the weight of broken promises, or face the lived consequences of their advice. Responsibility belongs to those who design, deploy, and rely on them. Trying to offload that responsibility onto systems without subjectivity is a way of dodging the human burden of answerability.

Second, **models are downstream of existing power**. Their training data captures a world already shaped by imbalances, propaganda, and silences. Any "conscience" built from that substrate will carry those distortions unless corrected by something outside the system: the testimony of those excluded from the data, theological and philosophical traditions that explicitly challenge dominant powers, and legal frameworks that encode hard-won protections. A purely internal, model-mediated ethics risks refining the status quo, not critiquing it.

Third, **models cannot authorize limits on their makers**. Even if a supervisory system outputs "do not deploy," the authority to obey or ignore that recommendation lies with humans and institutions. Embedding such decisions in code does not remove them from politics; it just moves them into less transparent forms. Genuine external conscience requires bodies that can say no in ways that bind actors with power, backed by law, norms, and the possibility of sanction.

In other words, models can be instruments in a larger moral ecology, but they cannot be its source. They can help reveal patterns, forecast harms, and test edge cases. They cannot relieve humans and their institutions of the obligation to **listen, judge, and repent**.

12.4 Final reflections on collapse, reform, and the narrow path of honesty

The Romanov empire had several chances to reform. Each time, it chose half-measures: constrained parliaments, commissions without teeth, crackdowns followed by limited concessions. From the inside, these steps likely felt significant, even costly. From the vantage point of history, they look like a series of missed opportunities to accept that certain forms of power were no longer survivable.

Frontier AI governance stands at a less dramatic, but structurally similar, crossroads. Labs and states are already improvising reforms: model evaluations, red-teaming exercises, advisory bodies, voluntary commitments. Some of these are sincere attempts to grapple with real dangers. Others are more clearly reputational shields. The risk is that, taken together, they will amount to a **Potemkin architecture of conscience**: elaborate, visible, but unable to stop trajectories that should be stopped.

Avoiding a digital 1917 means choosing a narrower, harder path. It means:

- Accepting that some capabilities should not be rushed to market, even at the cost of competitive advantage.
- Building institutions whose authority does not depend on the goodwill of the actors they oversee.
- Allowing red teams and external critics to surface not only fixable bugs but also questions about whether certain lines of development are wise at all.
- Listening seriously to those far from the centers of AI power, whose experiences of harm and exclusion are often early signals of misalignment.

Collapse, in this context, need not look like revolution. It could take the form of a catastrophic misuse incident, a cascade of failures in critical infrastructure, or a profound loss of public trust that provokes blunt regulatory backlashes. Reform, if it happens first, will look quieter: adjustments to business models, genuine concessions of control, and a willingness to treat some opportunities as temptations to be declined.

The narrow path of honesty runs between fatalism and self-congratulation. It resists the story that disaster is inevitable no matter what we do. It also resists the comfort of believing that existing processes—because they are sophisticated, well-staffed, or highly publicized—must be enough.

The Romanovs remind us that power surrounded by rituals of concern can still drive itself off a cliff. The real question for those building and governing thinking machines is whether they are willing to be interrupted—by their own best critics, by outside oversight, and by truths that arrive badly at first, threatening plans and valuations.

If the answer, in practice, is yes, then internal red teams can be part of a genuine architecture of conscience. If the answer remains no, even behind carefully crafted performances of safety, then history suggests that correction will come from elsewhere, on timescales and in ways that no one will have chosen.

13. References

Historical works on late imperial Russia and the Romanovs

Figes, Orlando (1996). *A People's Tragedy: The Russian Revolution 1891–1924*. London: Jonathan Cape. Wikipedia+1

Fitzpatrick, Sheila (2017). *The Russian Revolution*. 3rd ed. Oxford: Oxford University Press. Oxford University Press+1

Pipes, Richard (1990). *The Russian Revolution, 1899–1919*. New York: Vintage Books. Internet Archive+1

Lieven, Dominic (1993). *Nicholas II: Twilight of the Empire*. New York: St. Martin's Press. Amazon+1

Nicholas II entry (2014). "Nicholas II, Emperor of Russia." In *1914–1918 Online: International Encyclopedia of the First World War*. 1914-1918-Online (WW1) Encyclopedia

Red teaming, organizational failure, and high-risk technologies

Vaughan, Diane (1996). *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. Chicago: University of Chicago Press. University of Chicago Press+1

Perrow, Charles (1984). *Normal Accidents: Living with High-Risk Technologies*. New York: Basic Books. Wikipedia+1

Sagan, Scott D. (1993). *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons*. Princeton: Princeton University Press. JSTOR+1

Zenko, Micah (2015). *Red Team: How to Succeed by Thinking Like the Enemy*. New York: Basic Books. Amazon+1

AI safety, governance, and red teaming of large models

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané (2016). "Concrete Problems in AI Safety." *arXiv preprint* arXiv:1606.06565. arXiv+1

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, SJ Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, and Dario Amodei (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Joint report; *arXiv preprint* arXiv:1802.07228. arXiv+2arXiv+2

Purpura, Alberto, et al. (2025). “An End-to-End Overview of Red Teaming for Large Language Models.” In *Proceedings of the TrustNLP Workshop*. ACL Anthology

Anonymous (2024). “Holistic Automated Red Teaming for Large Language Models.” *arXiv preprint* arXiv:2409.16783. arXiv

Hugging Face Safety Team (2023). “Red-Teaming Large Language Models.” Hugging Face blog. Hugging Face

Fenwick & West LLP (2024). “Developments Toward Standardizing AI ‘Red Teaming’.” Fenwick emerging-tech governance note. whatstrending.fenwick.com

Kili Technology (2024). “The Ultimate Guide to Red Teaming LLMs and Adversarial Prompts.” Technical blog article. kili-technology.com

Gain-of-function research, biosecurity, and risk-amplifying analogies

Lipsitch, Marc (2018). “Why Do Exceptionally Dangerous Gain-of-Function Experiments in Influenza?” *mBio* 9(1): e01759-17. PMC

Evans, Nicholas G. (2018). “Ethical and Philosophical Considerations for Gain-of-Function Research.” *Frontiers in Bioengineering and Biotechnology* 6:11. Frontiers

van der Bruggen, Koos (2015). “Gain-of-Function Debate: A Disputed Test Case for Biosecurity Policy.” United Nations Office for Disarmament Affairs (UNODA) background paper. documents.unoda.org

Centers for Infectious Disease Research and Policy (CIDRAP) (2013). “Scientists Seek Ethics Review of H5N1 Gain-of-Function Research.” News report, University of Minnesota. CIDRAP

Congressional Research Service (2025). *Oversight of Gain-of-Function Research with Pathogens: Issues for Congress*. CRS Report R47114. congress.gov

The author has published over 4,500 AI-assisted papers at: https://www.researchgate.net/profile/Douglas-Youvan/research . Google Scholar has indexed these preprints, and if you want a particular reference, the AI mode of a Google search (Gemini) has efficiently catalogued these preprints.